MODEL IDENTIFICATION IN THE BIOCHEMICAL SYSTEMS THEORY

SRIDHARAN SRINATH

(M. Tech, IIT Roorkee, India)

(B.Tech, Anna University, India)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Department of Chemical and Biomolecular Engineering

National University of Singapore

2012

ACKNOWLEDGEMENTS

It gives me great pleasure to express my profound sense of gratitude to my PhD advisor, Dr. Rudiyanto Gunawan, for his encouragement and meticulous guidance during the course of my graduate studies. I also would like to appreciate his patience especially when I was stuck and when progress was slower. Without his support it would have been impossible for me to come this far. He is constantly willing to share his experiences and I have learnt innumerable lessons and insights on research from the discussion sessions with him.

I am also thankful to Dr. Saif A Khan, for all his support in culmination and submission of this thesis. I am also indebted to my PhD examiners, Dr. Rajagopalan Srinivasan and Prof. Rangaiah. G.P. for their acceptance of being in my examination committee and their constructive comments during my qualifying examination which helped in moulding this thesis into a better shape.

My words fail to express my sincere thanks to my lab colleagues, Suresh Kumar Poovathingal, Thanneer Malai Perumal, Lakshminarayanan Lakshmanan, Jia Gengjie and Tam Zhi Yang for providing a congenial environment in lab and all the healthy discussions during group meetings. Their critical feedback and comments helped me in bringing about a good thesis. The lunch and tea-time discussions with Thanneer, in particular, have helped me discuss and brainstorm new ideas about systems biology, systems engineering and research in general. Many of these discussions with him have been influential in developing the framework of this dissertation.

I appreciate Department of Chemical and Biomolecular Engineering (ChBE), NUS, for providing me the infrastructural and necessary facilities to carry out this research work and Ministry of Education for providing financial support for my PhD. My heartfelt gratitude to the Professors of ChBE for the valuable lectures and seminars during my coursework and PhD duration. In particular the course on Mathematical modeling of chemical and environmental engineering handled by Dr. Lakshminarayanan Samavedham was instrumental in me developing the fundamental concepts of mathematical modeling and MATLAB. His organization of materials and methodology of teaching will always be a source of inspiration for me. Finally, I also offer my thanks to Mr. Boey Kok Hong and Ms. Samantha Fam for their unrelenting technical and administrative help.

I've made quite a few friends in and around NUS to keep myself sane. I would like to thank all of my friends who have been a part in some way or other during my PhD journey.

Last, but most importantly I highly appreciate the cooperation and affection showed by my parents, wife and my sister who have been a constant source of inspiration, patience and moral encouragement to me.

Sridharan Srinath

CONTENTS

ABSTRACT

Recent advances in technology permit high throughput experiments at genomic, transcriptomic, proteomic and metabolomic levels. The information obtained from time-series data, however, is implicit and requires extensive data analysis. Mathematical modeling of biological systems has found increasing applications for investigating dynamics in complex cellular processes and has given rise to a new field called *systems biology*. In systems biology, biological processes like signal transduction and metabolic networks are often modeled using differential equations. These models often include many unknown parameters like enzymatic reaction rate constants, which are to be determined by fitting to time-course experimental data. Using the power-law formalism, the Biochemical Systems Theory (BST) coupled with high-throughput biological measurements transform the model identification into an inverse problem of estimating model parameters from experimental data.

Given time-series data and a model, parameter estimation can be thought as the "inverse problem" of generating predictions from model. Despite the large number of publications on this topic, this task remains the bottleneck in the application of BST modeling in biologically related area. Many studies in the literature have focused on developing comprehensive parameter estimation techniques that exploit many of the mathematical features of canonical models within the BST, such as S-systems or generalized mass action (GMA) or linlog models. However, many challenges arise from the same underlying problem; incomplete and noisy measurements lack the necessary information in order to accurately estimate the model parameters. This is a parameter

identifiability problem. Thus, the focus of this work is to investigate the identifiability of metabolic network models, and to suggest model refinement or experimental design that maximizes the number of estimable parameters from data.

Two types of identifiability property are considered. First, *a priori* identifiability analysis yields the identifiable parameters under the assumption of noise-free data. Parametric sensitivities are used as a basis for selecting the *a priori* identifiable parameters. Secondly, practical identifiability gives the identifiable parameters when the data are contaminated with noise. In other words, this analysis gives the accuracy with which the parameters can be estimated. The practical identifiability analysis methods are based on linear(ized) and nonlinear regression analysis, particularly the statistical inference of confidence interval or region of the parameter estimates. The applications to two inverse modeling problems within the BST point to the lack of parametric identifiability as the root cause of the difficulty faced in the inverse modeling. Although this work focuses on the BST models, the analyses can be applied to other types of models, and the issue of parameter identifiability is expected to be a common problem in other biological modeling.

Motivated by the results of the analysis to S-systems and GMA models, we developed methods based on nonlinear regression to test the identifiability of decoupled and linlog systems. The problem often faced in the parameter estimation of these models is the difficulty in integration due to numerical stiffness, constituting almost 95% of time spent for the parameter searches [1]. Two alternative BST formalism considered are: Decoupled Models and Linlog models. Numerical integration of such dynamic models can be circumvented by fitting the differentials with slopes that are estimated from the

time-series data at all measured points, essentially decoupling the ODE model [2]. This work focuses on the parameter identifiability analysis of such decoupled systems. The treatment of noise in the data was also taken into account in detail. The analysis was applied to the decoupled versions of two previously published power-law models of metabolic networks: glycolytic pathway in *L. lactis* and recombinant *E. coli* growth. The results were then compared with that of the parameter estimation of the original ODE models, revealing the differences between decoupled model and its corresponding BST model. Linlog model is a new mathematical framework combining a general kinetic model and theorems from MCA. The number of parameters is minimal and all rate equations have the same mathematical structure as BST models. Parameter estimation of BST models is a bottleneck and hence linlog models are a good alternative for the BST models as it ameliorates the parameter estimation task. The parameter identifiability of linlog models were performed using the methods developed previously. The results highlighted the fundamental problem of linlog models: rate being undefined at zero concentration. As a result of this problem, parameter identifiability of the linlog models was poor.

The next work deals with Design of Experiments (DOE). Dynamic mechanistic models allow a better understanding of the various phenomena taking place within the system of investigation. These models usually involve unknown parameters which are often estimated by calibrating the model with some experimental data. The collection of these experimental data is costly and requires precise planning of experiments that would give maximum information at minimum resource utilization. Model-based design of experiment (MBDOE) offers an avenue for combining modeling and experimental efforts

such that knowledge generated from past experimental data, captured by the model equations and parameters in the present iteration, is used to guide subsequent experiments. To this end, a multiobjective optimization design criterion that accounts for curvature effects was developed which resulted in better performance than the conventional FIM-based designs.

The final goal of this PhD thesis was to integrate the tasks of identifiability analysis, DOE and parameter estimation together into a useful MATLAB tool. The main focus was on integration of the methods developed previously for identifiability analysis and DOE in an iterative loop along with parameter estimation for which a two-phase strategy is adopted to overcome the stiffness problem. This iterative procedure was applied to a five variable gene regulatory network and the results suggested that the iterative method helps in obtaining a model with higher number of identifiable parameters than the original model and less parameter errors.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Biology can be thought as a study of self-replicating chemical processes. Advances in molecular biology have permitted high throughput measurements at multiple scales in a cell, from genomic to transcriptomic, proteomic and metabolomic. With the wealth of such data, one fact that becomes immediately clear is that the typical biological system is complex and that phenotype is an emergent system behavior from interactions among biological components. These revelations have changed the perspective on how biological research is or should be carried in the post-genomic era, giving birth to systems biology. The field of systems biology combines methods and concepts from systems engineering, computer science, statistics and biology, among other things, with the aim to study and understand biology from the systems level [3].

One of the cornerstones of systems biology is the reconstruction of the complex network of biological interactions in the form of a mathematical model [4, 5]. The continued advances in experimental techniques to obtain single cell measurements further provide an opportunity for developing detailed mechanistic models of many biochemical networks, from signal transduction pathways to metabolic pathways. The complete reverse engineering of a biochemical network in general needs to reconstitute both the topology and kinetics (rules) of interactions. The mathematical challenges in this problem range from selecting the most appropriate model structure to identifying of model parameters from noisy measurements.

There is no absolute answer to the question "what is the best or exact mathematical representation of a biochemical reaction/network?" In fact, an exact model does not exist and George Box has put this so eloquently, stating that "all models are wrong, but some are useful" [6]. That is, mathematical models are only abstractions of the real process. The real task of building a model is to construct such an approximation that is accurate enough to be useful and simple enough to be mathematical or computationally tractable. However, predictive mechanistic models of biochemical networks require sufficiently detailed knowledge about the system. As model predictions often depend on the value of model parameters, accurate and reliable quantification of the parameters is essential. However, for most biological processes, the system parameters often cannot be directly measured *in vivo* and thus have to be estimated from quantitative information on reaction rates and/or molecular concentrations. For these reasons, parameter estimation is an important part of systems biology [7-9].

This chapter is organized as follows. First, an introduction to mathematical modeling in biology is presented, wherein the history and evolution of the use of mathematical models in biology is reviewed. Then the concept of model identification in biology is discussed, followed by the short descriptions on the various models used in systems biology. Afterward, the focus shifts specifically to challenges in the parameter estimation of power-law models, including identifiability analysis and design of experiments. Finally, an integration of these components into iterative model identification cycle is proposed and presented. This chapter is terminated with the organization for the rest of the thesis.

1.1 MATHEMATICAL MODELING IN BIOLOGY: A HISTORICAL PERSPECTIVE

The success of biology in the 18th, 19th and 20th centuries relied much on reductionism approach, where in order to understand biological functions at the organism-level, it is essential to understand the organs and to understand organs, it is then necessary to study tissues, cells and other sub cellular components. Such approach is hoped to provide the complete knowledge of fundamental elements and processes and once these elementary units are known, the cells and organisms can be rebuilt bottom-up. So, the analysis of a whole organism or organ function has been reduced into the study of specific physiological processes and characterization of biochemical pathways, which constitute a series of chemical reactions occurring within a cell. The reductionist approach has led to the modern day experimentation in life sciences and also provides the rationale for the endeavors like Human Genome Project [10]. But, as Henri Poincaré once said "the aim of science is not things in themselves but the relations between things", even after having the knowledge of almost all constituents of cells, it was not possible to reliably predict how an organism would respond to new conditions or stimuli [10]. The growing amount of experimental data and experience with microorganisms and cell lines did not help in the true understanding of biological phenomena. Gradually it was apparent that a paradigm shift was utmost necessary.

Mathematical modeling has become an indispensible tool in biotechnology and biological studies with myriad applications from metabolic engineering to cancer therapy. But, the use of mathematics to describe biological phenomena only started in 1943, when Erwin Schrödinger gave a series of talks in Dublin on the topic "*What is life?"* The central idea of his talks was that biological systems follow physical and not metaphysical

laws and that they can be described by mathematical equations [11]. In 1952, Hodgkin and Huxley explained and underlined their experimental data with a mathematical model, which was critical in understanding how neurons function [12]. A few years later, Denis Noble elaborated and modified this model to obtain the first mathematical model of the heart [13]. The Hodgkin-Huxley model and its variants are still a crucial part of computational neuroscience and are being used today by researchers across the world. The success story of Hodgkin-Huxley model is attributed to the fact that they were able to estimate the model parameters from their experimental data. But in areas like cellular signaling, gene regulation and metabolic pathways, model identification is still very challenging [14].

In the late 1940s, several researchers put forth novel concepts simultaneously. Some of the notable works like Norbert Wiener's cybernetics [15], Turing's ideas of automata and computability [16], von Neumann and Morgenstern's game theory [17] and Shannon's information theory [18] were popular in the areas of physical sciences. In the same period, mathematicians and engineers pushed for the development of systems theory. Wolkenhauer [19] pointed out that one of most important contributions of the systems biologist Robert Rosen in the 1960s may have been his introduction of biological theory sufficient to investigate final causation without implying teleology [20]. During the same times, few researchers like Rashevsky [21], Ashby [22] and Rosen [23] proposed applying systems theory concepts to comprehend biological phenomena. Goodwin [24], Heinmets [25] and others also developed kinetic models for gene regulation describing induction and repression, thereby recapitulating in mathematical terms the ideas of Jacob and Monod [26] on the functioning and regulation of operons in

bacteria. As a consequence of growing need of more quantitative approach in biology, *Journal of Theoretical Biology* was started in 1961. The same period also saw the birth of Biochemical Systems Theory (BST) [27, 28] in late 1969 and Metabolic Control Analysis (MCA) [29] in 1973. These two frameworks have parallel histories, and theories developed for each framework converged in the last 15 years [30, 31].

In summary, the majority of the ideas of today's systems biology emerged during the 1960s. But the acceptance of the systems theory concepts in biology was not smooth and it took a long time before biologists were convinced the advantages of using mathematical models. The reductionist concept was prevailing among the researchers throughout the 20th century [10]. The concept of reductionism was a huge success and so it begs the question why should biologists look beyond reductionism. The answer is actually two-fold.

a. The research conducted using reductionist approach is generating so much data that it is becoming mandatory not only to collect and store these data in databases, but also to develop a functional context within which each experimental result becomes meaningful. Without a functional context, data are just simple description of the features of the system in hand and only their integration within their physical, spatial and functional surroundings yields insights and ultimately knowledge and understanding.
b. The main shortcoming of reductionist approach is the observation that components of a biological system behave differently *in vivo* than in isolation [10]. The difference might be unimportant in some cases, but in certain cases it might drastically affect the function of the particular component [10].

1.2 MODEL IDENTIFICATION IN BIOLOGY

The construction of a model is often referred to as model identification, which is a well-established field within systems engineering [32-34]. In [34], Ljung has defined the terminologies related to system identification, in which two are of particular interest: modeling and identification.

Modeling: "A study is made of the mechanisms inside the system, and from basic physical (biological, economical, etc.) laws and relationships a model is inferred".

Identification: "Data are collected from the system and a model is selected that describes these data sequences well. This procedure could very well be combined with a modeling step to determine the numerical values of certain physical constants in the model".

Successes in the modeling of biological systems have provided good insights into the physiological properties of an organism [10]. Among these, metabolic pathways, comprising a set of enzymatic reactions involved in the cellular metabolisms, have generated a lot of interest because of their industrial and biotechnological relevance, e.g. in the microbial production of penicillin or biofuels. These models are essential to comprehend and predict the effect of changing of enzyme activities on the desired flux and metabolite levels, and to guide the metabolic engineering efforts [35].

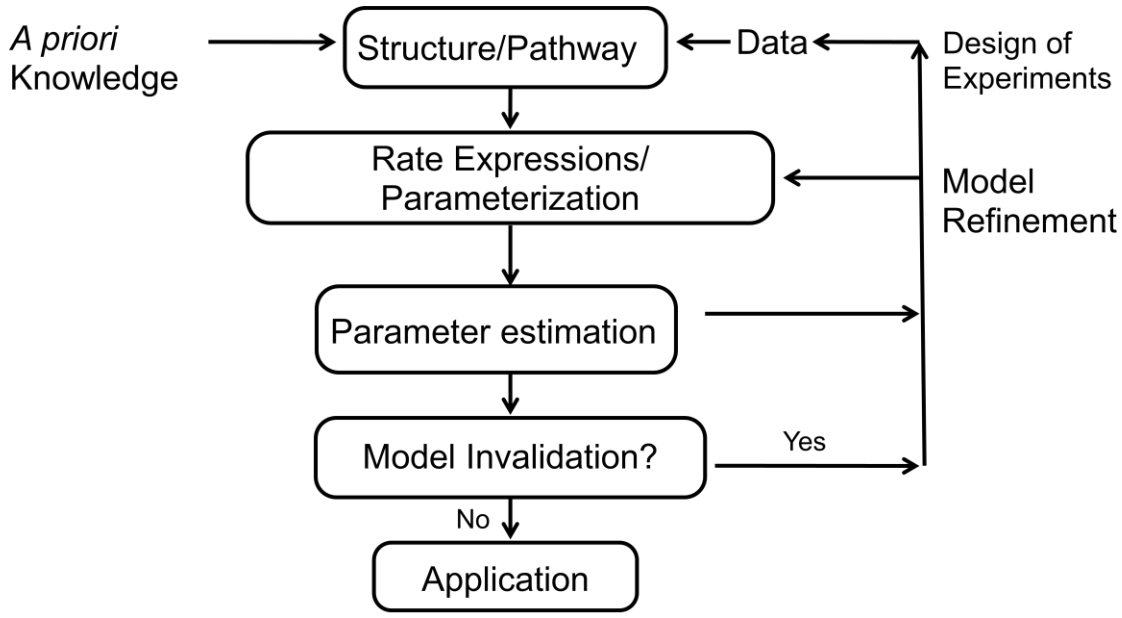

Figure 1.1. An Iterative Procedure of Model Identification

The procedure of model construction is typically iterative, in which wet-lab experiments generate the biological observations and data needed for model formulation and identification, while *in silico* simulations are used to (in)validate models and design the most informative experiments [36, 37]. In general, the iterative steps include (see Figure 1.1);

- formulation of network structure,
- selection of kinetic equations,
- assignment and/or estimation of model parameters,
- validation of the model.

The model identification of biological systems provides great interest and major challenges to systems biologist. The complexity of a typical biological system limits the applicability of methods which have been devised for the identification of complex

engineering systems due to several reasons that will be discussed below. In this thesis, I will focus on the modeling of metabolic pathways.

Based on prior knowledge of a biological pathway, the first step of the iterative procedure is the development of hypothesis regarding the biological network structure. In this step, knowledge or assumption regarding relationships among components in the system is used to create a network graph, where the nodes represent the biological molecules (e.g. metabolites) and the arrows represent the (enzymatic) transformation or regulatory actions among them. Following this, the graph is then converted into a dynamic model by prescribing the rules or kinetics governing each arrow using mathematical equations. For a dynamical system, these models typically are formulated as a set of ordinary differential equations. Once the model equations have been proposed, the next step is to assign or if not known, to estimate the model parameter values that make the model consistent with the experimental data. With a full set of parameter values, the model is finally validated against independent experimental observations. Once validated, the model can then be used for many applications such as metabolic engineering.

The model identification procedure may not necessarily nor should it be expected to lead to a single, unique solution. In fact, there could be many candidate models that can reproduce the experimental observations. This is known as incomplete model identifiability and the issue depends on many factors, such as

- Model structure and parameterization
- Experiment design

- Quality and Quantity of data generated
- Variables measured

In this thesis, focus will be on a subset of this problem, which deals with the identifiability of model parameters.

1.2.1 Peculiarities of a Biological System

Although mathematical modeling has found wide applications in science and engineering, due to the peculiarities of the biological system, many of the existing methods need to be modified to meet the demands of biological system identification. Carson and Finkelstein [38] have pointed out the special features of biological systems in relation to the identification process. The peculiarities of biological system are described in five aspects in [10]. First, biological processes are highly non-linear and complex. For example, a simple yeast cell has over 6000 genes, while humans have about 100,000 proteins. So, the model complexity arises from the large number of biological components and even larger potential interactions among them. Thus the model proposed must be able to capture these non-linear aspects and should be able to handle the large-scale nature of the system. Second, in biological systems, dynamic and continuous change of the species is often of interest. Hence, the models must be time-dependent, which in most cases is modeled using a set of differential equations. The third is the scaling issue. As biological system is composed of various levels of components (genes, proteins and metabolites) and interactions, the scaling of the proposed model should be able to handle these different levels of interactions that could span several orders of magnitude in length and time scales. The fourth is the inherent stochastic behavior of the biological systems, especially when only very few molecules are involved in the process.

So, the proposed model should be able to capture this stochastic behavior in addition to the deterministic behavior. Finally, biological reactions rarely occur in a homogeneous environment, but are segregated to organelles or compartments. Although an ideal model for a cell should capture all these peculiarities, unfortunately, no existing modeling framework can meet all of these requirements.

In practice, the modeling framework can be chosen by considering the importance of the following four properties in the system:

- dynamic or static
- continuous or discrete
- deterministic or stochastic
- spatial or homogeneous

Metabolic pathways are usually modeled using a continuous deterministic and homogeneous model. Stochasticity is usually more prominent in gene regulatory networks, and if the spatial dependence is not important, then the cellular environment can be assumed to be homogeneous. Since the transient behavior is of interest, as captured by time-series data of metabolic pathways, the focus of this thesis will be on deterministic and dynamic models of metabolic networks.

The issues prevailing in modeling of biological systems can be categorized into four areas, as discussed below.

1. **Data related issue:** The biological data are usually noisy due to measurement errors and are seldom complete and this noise in the data could corrupt the actual information in it. Missing data problem is quite common wherein either the data

are sparsely missing or data collection is lacking at certain time points or the entire time-series is missing. Although there are methods like Jia *et al* [39] to handle the missing data, sometimes this problem leads to ill-conditioned data matrix, which may be due to the collinearity among the time series data or non-informative data.

2. **Model related issue:** The selection of a model could influence the parameter estimation problem to certain extent. The various modeling frameworks are discussed in the following section. Some key pointers in choosing a model are the ability to capture the dynamics of the time profile, mathematical simplicity and tractability and interpretability of the results within the biological realm. Chou and Voit [1] discussed in detail the pros and cons of various modeling frameworks used in this field. At least in this thesis this issue is not a problem since the focus was towards BST modeling framework.
3. **Computational issue:** This issue is really challenging and has been a prime focus in many articles published in inverse modeling. The peculiarities of biological systems, discussed above, clearly indicate the biological models potentially contain many components; the systems are usually nonlinear and are formulated as a set of ODE to capture the transient behavior. So, the computational issues are emerge out of this are the lack of convergence towards the global optima, significant time spent in integration of the ODE and improving the computational efficiency of the parameter estimation process as a whole.
4. **Mathematical issue:** The last source of problem that hinders the parameter estimation task is the mathematical redundancy in the models. These include that

different sets of parameter values fitting the experimental data equally well (parameter identifiability) or non-equivalent solutions exhibit similar residual errors. The mathematical issue was tackled in this thesis.

1.3 MODELING FRAMEWORKS

The first and foremost step in any modeling process is to choose a particular modeling framework depending upon the characteristics of the system and the desired objective. Biological models can vary from simple algebraic models to state-space models to a stochastic chemical master equation (CME). In order to model a dynamic process, ordinary differential equations are often used, while steady-state models can be derived from these equations by setting the time derivative to zero to give algebraic equations. If the process in hand is intrinsically stochastic, then a CME is often used to model such system. The most commonly used models to describe metabolic pathways are kinetic models and stoichiometric models. In addition, for kinetic models, power-law formalism under the biochemical systems theory (BST) framework is also often used as canonical model equations. The BST models are derived by using the power-law formalism for the reaction fluxes, such that the couplings among the metabolites or states (i.e. model structure) can be inferred from the values of the parameters, as will be detailed below. The two most popularly used power-law models in the BST are the S-systems and Generalized Mass Action (GMA) models. In the following sections, the two common types of models used for metabolic pathways – stoichiometric and kinetic, are discussed.

1.3.1 STOICHIOMETRIC MODELS

Mathematical models of metabolic pathways are built from stoichiometric and kinetic information of the enzymatic transformations. The stoichiometry is time invariant and the stoichiometric matrix essentially describes the metabolite transformations (i.e. mole balances), where the rows and columns correspond to the metabolites (nodes) and the reaction (arrows), respectively. The sign of the stoichiometric coefficient indicates whether the particular reaction increases (positive) or decreases (negative) the concentration of a particular metabolite (corresponding to the row). The value indicates the stoichiometric relationship and should be an integer. A value of zero indicates that the metabolite and the reaction are not related. Stoichiometric models consider the product of the stoichiometric matrix **N** and a vector of metabolic fluxes v, to describe the dynamics of the metabolite concentrations X using a set of ordinary differential equations, given below: [40, 41]

$$\frac{dX}{dt} = \mathbf{N}v. \tag{1.1}$$

The Flux Balance Analysis (FBA) is a prominently used analysis to obtain the flux distribution at steady state. The fundamental principle underlying FBA is the conservation of mass. A flux balance is written for each metabolite (X_i) to yield the dynamic mass balance equation as written above in Eq. (1.1). The main use of pure stoichiometric models is to determine the fluxes v in the network under the steady state assumption, given below.

$$\mathbf{N}v = 0 \tag{1.2}$$

This steady state equation is solved for the fluxes to obtain the flux distribution. Typically, the number of metabolic fluxes (v) is greater than the number of metabolites (X), resulting in a plurality of feasible flux distributions. This range of solutions is indicative of the flexibility in the flux distributions that can be achieved with a given set of metabolic reactions. The basic premise of FBA is that the cells evolve with a specific objective function, most commonly, maximizing cell growth. In FBA, the solution to Eq. (1.2) is formulated as a linear programming problem, in which the flux distribution is found by optimizing a particular objective [42].

Flux balance models are simple and easy to build, but their main drawback is their often limited predictive power [1], due to the lack of dynamical and regulatory information in the model formulation. So, in this thesis, the focus is on kinetic models, rather than the pure stoichiometric models.

1.3.2 KINETIC MODELS

When information about the kinetics of specific metabolic transformation is available, it is then possible to model the dynamics of these processes in Eq. (1.1) [43]. The functional forms of kinetic models are usually based on the law of mass action, Michaelis-Menten enzyme kinetics, or other canonical equations, e.g. power law. In this thesis, power-law based equations within the BST are considered. Such canonical models are useful since the structure has been parameterized.

There most popular formalism can be grouped into two categories, power-law equations [44] and linear-logarithmic (linlog) equations [45]. In power-law models, rates and variables are linearized in logarithmic spaces, and the models include the

Generalized Mass Action (GMA) and S-systems models. On the other hand, a linlog model is a hybrid linearization model, in which the rates are in linear terms, but the concentrations are in logarithmic scale. There are two possible sub-types of linear-logarithmic models: (log)linear and linlog. While the power-law models are used within the BST, the linlog model is often used in the Metabolic Control Analysis (MCA) [46]. The theory underlying the analysis in the BST and MCA, though different, have overlapping interpretations, and a unifying theory has been proposed to merge the two analyses [30]. Specifically, Marin-Sanguino *et al* [47] developed a common framework for four most commonly used canonical models: two power-law and two linlog models described above, within the BST. In the following sections, the canonical models of the BST will be detailed further.

1.3.2.1 S-systems

A widely used canonical model of metabolic networks is the S-systems. The model consists of a set of differential equations, which can be written as

$$\dot{X}_i = V(X_1, X_2, ..., X_{n+m}). \tag{1.3}$$

where the right hand side V depends on n dependent variables $X_1, X_2, ..., X_n$ and m independent variables $X_{n+1}, ..., X_{n+m}$. The dependent variables in an S-system can be

- metabolites in a metabolic pathway model,
- concentrations of activated and non-activated proteins in a signal transduction network model, or
- levels of expression in a gene regulatory network model.

The multivariate function V can be split into two parts, V_i^+ and V_i^-, where the former denotes production and the latter denotes degradation process, leading to the following equation:

$$\dot{X}_i = V_i^+(X_1, X_2, ..., X_{n+m}) - V_i^-(X_1, X_2, ..., X_{n+m}). \quad (1.4)$$

The flux V_i^+ in the above equation is approximated using a power-law, which leads to

$$V_i^+ \approx \alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}}. \quad (1.5)$$

A similar expression for V_i^- can be obtained as

$$V_i^- \approx \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}}. \quad (1.6)$$

Substituting equations (1.5) and (1.6) in (1.4), yields

$$\dot{X}_i = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}}. \quad (1.7)$$

The ODE above is called an *S-system* model, where *S* refers to *synergism* and *saturation*. The parameters in the approximation are the rate constants α_i and β_i, and the kinetic orders g_{ij} and h_{ij}. The rate constants can take any value in the set of positive real numbers, but the kinetic order can be any real value. The sign of the kinetic order parameters also has a physical interpretation, in which a positive value indicates substrate or activation and a negative implies inhibition [44]. So, the structure of a pathway, i.e. the

connections among the variables (X_i) and the directions of these connections, can be directly identified from the parameter values.

1.3.2.2 Generalized Mass Action (GMA)

Unlike in the S-systems, the GMA formalism does not use aggregated influxes and effluxes terms. Instead, each reaction is approximated by a power-law equation, giving

$$\dot{X}_i = \sum_{k=1}^{n+m}\left(\pm\gamma_{kp}\prod_{j=1}^{n}X_j^{f_{ijk}}\right). \tag{1.8}$$

where the rate constants γ_{ik} are again non-negative and the kinetic orders and f_{ikj} can take any real value. The GMA and S-system differ only at the branch points (i.e. when there is more than one arrow going into or out of a node).

There are many advantages and disadvantages of the BST formulations and these have been discussed at length elsewhere [27, 28, 44]. To summarize, some of these are:

- richness in the structure which is able to capture all forms of non-linear behavior, e.g. oscillation, chaos;
- the BST models can be set up without the mechanistic information of the system;
- the highly ordered mathematical structure facilitates easy mathematical and numerical analyses.
- but, the number of parameters increases rapidly with the number of metabolites [48], often leading to over parameterization of the model

Power-law models have been initially used to model metabolic pathways, but they have also been applied to other levels of biological systems, including genetic networks [49], multi-level systems [50] and cell signaling [51].

1.3.2.3 Linlog formalism

Another approximate linear-logarithmic (linlog) kinetic model was introduced by Hatzimanikatis [52] and expanded by Visser and Heijnen [53, 54], which has been shown to have a good approximation quality, standardized and relatively fewer parameters. This particular formalism was derived by combining a general kinetic model and theorems from the Metabolic Control Analysis (MCA). The model includes general expressions giving steady-state fluxes and metabolite levels as a function of enzyme levels, extracellular concentrations and the control and response coefficients. An overview of different approximation of enzyme kinetics are presented in [55]. In particular, the rate is proportional to the enzyme concentration (e), linear in the logarithm of the dependent (X), and independent (C) metabolite concentrations (hence non-linear in metabolite concentration) and linear in kinetic parameters (p_i and q_i), i.e.

$$v = e\left(a + \sum p_i \ln X_i + \sum q_j \ln C_j\right). \tag{1.9}$$

Generally in MCA, it is preferred to normalize the kinetics with a reference such that the rate is defined by the reference elasticities $\left(\varepsilon^0\right)$ and the reference parameters: flux$\left(j^0\right)$, enzyme concentrations $\left(e^0\right)$, and metabolite concentrations$\left(x^0, c^0\right)$. For *in vivo* kinetic experiments, the reference is taken at the steady-state condition before the perturbation is applied. Using reference values, Eq. (1.9) becomes;

$$\frac{\mathbf{v}}{\mathbf{v}^0} = \left[\frac{\mathbf{e}}{\mathbf{e}^0}\right]\left(i + E_X^0 \ln\frac{\mathbf{X}}{\mathbf{X}^0} + E_C^0 \ln\frac{\mathbf{c}}{\mathbf{c}^0}\right). \quad (1.10)$$

in which,

- E_x^0 and E_L^0 are matrices containing the elasticity coefficients for dependent and independent metabolites.
- i is the vector of ones
- $\mathbf{X}/\mathbf{X}^0$ is a vector of relative dependent metabolic concentrations
- $\mathbf{c}/\mathbf{c}^0$ is a vector of relative independent metabolic concentrations
- $\mathbf{e}/\mathbf{e}^0$ is a square diagonal matrix containing relative enzyme levels

The linlog kinetics is gaining popularity rapidly and has been used for metabolic network modeling and parameter estimation [56-58]. Recently, Marin-Sanguino *et al* provided a common framework for S-system, GMA and linlog models and compared two philosophies: the application of the design equation and the solution of constrained optimization problems [47]. The linlog models are discussed at a greater detail in Chapter 3.

1.3.2.4 Comparison of Canonical Models

In the previous sub-sections, three different canonical models were introduced. The obvious follow-up question would be, "when to choose which formalism?" The answer is dependent on the particular problem and the information available. For instance, the GMA model, as mentioned earlier, is in essence a stoichiometric model with added kinetic information through power-law approximation. However, the GMA

models cannot be used for steady-state calculations by simple algebraic manipulation, unlike the S-systems. This is because S-systems having a simpler structure (only one production and one consumption term in the RHS) can be reduced to a set of algebraic equations in logarithmic space at steady state. Although a linlog model also permits easy steady-state calculations, it cannot represent certain non-linear behaviors as the structure is basically linear [59]. Also the linlog models can become erroneous when the substrate concentrations are close to zero [56, 57]. Nevertheless, both the power law and linlog approximations are local approximations and can perform well as long as the variables do not vary too much. One unique feature of all canonical models is that their parameters can be mapped almost uniquely onto the structure of a pathway. This conclusion shifts the heavy burden of identifying the structure of a pathway onto the estimation of parameter values. This is the motivation behind choosing canonical models ahead of other non-canonical models (like law of mass action or Michaelis-Menten enzyme kinetics) in this thesis.

1.4 PARAMETER ESTIMATION

After the kinetic modeling framework has been decided (S-system or GMA or linlog), the next step is to assign and/or estimate the parameters of the model. But, many of model parameters are usually unknown *a priori*, and thus they are usually estimated from data. Despite the advances in the parameter estimation field, existing techniques often fail to yield accurate parameter estimates in biological systems. There have been many in-depth studies regarding the parameter estimation of non-linear biochemical systems, e.g. Kimura *et al* and Voit *et al* highlighted many foreseen and unforeseen

challenges in parameter estimation due to measurement noise and the explosion in the number of parameters [49, 60]. Parameter estimation step is arguably the bottleneck of mathematical modeling of biochemical systems [14].

In the following section the various methods developed for parameter estimation are reviewed: forward (bottom-up), inverse (top-down) and steady state estimation. In classical bottom-up approach, the system structure and rate laws for every reaction were integrated into a dynamical model and instantiated with known parameter values from the literature. In the top-down approach, the biochemical network topology and the parameter values are inferred directly from observed time-series data. When the network topology is known, the top-down strategy is equivalent to a constrained parameter estimation task.

1.4.1 Forward or Bottom-up modeling

This is the traditional reductionist approach of mathematical modeling in which, metabolic models are developed using "local" kinetic information. For example, a purified enzyme is studied and characterized to determine the optimal temperature, pH and other cofactors, modulators and secondary substrates. Since the metabolic flux depends on the enzyme and substrate concentrations, the above mentioned information is then converted into a mathematical rate expression. After acquiring sufficient information about each rate expression, a model that that integrates all the relevant expressions into a comprehensive mathematical model is built. If the assumed network structure reasonably approximates the true pathway, the model should simulate the behavior of the network, at least qualitatively if not quantitatively [2, 14, 61]. There are many recent studies that were based on this approach of modeling, e.g. a fermentation model of *Saccharomyces*

cerevisiae [30, 62], citric acid cycle model of *Aspergillus niger* [63, 64] and purine metabolism [65].

1.4.2 Inverse or Top-down modeling

High-throughput experiments at genomic, transcriptomic, proteomic and metabolomic levels can aid overcoming the above mentioned challenges by providing "global" kinetic data. The experimental tools which permit the generation of dynamic metabolite concentration profiles include Nuclear Magnetic Resonance (NMR), Mass Spectrometry (MS), High Performance Liquid Chromatography (HPLC) and flow cytometry [2]. The data are obtained under the same experimental conditions, within the same species and sometimes *in vivo*. These measurements contain information regarding the functional connectivity and regulation of the biological network. The information within time-series data, however, is implicit.

The construction of models that describe the experimental data relies on the estimation of unknown, possibly case-specific, model parameters by way of experimental data fitting, also known as inverse modeling. The most challenging part of the modeling process using BST models is the parameter estimation step [1], which has attracted considerable efforts from many scientists. In general, the parameter estimation is usually started with an assumption that the topology of the network and the proposed model are correct, though the use of power law is already an approximation of the reality. Unfortunately, the true topology of biological system is often not completely known. In such cases the inference of network topology becomes a bottleneck, more difficult than the parameter estimation step [1].

In the BST, the parameter estimation can also be viewed as structure identification because of the one-to-one relationship between structure and parameters. There have been many recently developed parameter estimation techniques for BST models, review of which can be found in [1]. Some of these methods are applicable to other models, but many of them take the advantage of the specific structure of power-law models. These methods can be classified as:

a) Methods based on integrating the differential equations
b) Methods based on slope estimation
c) Methods based on constraining the parameter search space

A detailed review of these methods can be found elsewhere [1]. However, even after more than 100 publications in this area (Table 1.1), the estimation of BST parameters from data is still an open problem, in which different estimation techniques will often give different parameter estimates. This is a tell-tale sign of parameter identifiability issue, which will be discussed briefly in the next section and in greater detail in the next chapter.

Table 1.1 Comparison of various algorithms used for parameter estimation in BST models **[1]**

S.No	Authors	Year	Method	Model	Case study used
1	Kikuchi *et al* [66]	2003	GA	S-system	(a)
2	Voit & Almeida [2]	2004	• Decoupling • ANN	S-system	(b)
3	Kimura *et al* [49]	2004	Decomposition method	S-system	(a) (c)
4	Tsai & Wang [67]	2005	• Modified collocation • Decoupling	S-system	(a) (d)
5	Marino & Voit [48]	2006	• Decoupling • Gradient-based	S-system	(b)
6	Kim *et al* [68]	2006	Genetic programming	S-system	(b)
7	Tucker & Moulton [69]	2006	Interval analysis	S-system	(a) (b) (f)
8	Polisetty *et al* [70]	2006	Branch and bound	GMA	(g) (h)
9	Gonzalez *et al* [71]	2007	Simulated Annealing	S-system	(b)
10	Kutalik *et al* [72]	2007	Newton-flow analysis	S-system	(b) (c)
11	Tucker *et al* [69]	2007	Constraint-propagation	S-system GMA	(b) (i)
12	Marin-Sanguino *et al* [73]	2007	Geometric programming	GMA	(h)
13	Liu & Wang [74]	2008	HDE	S-System	(a) (c) (i) (j)

(a) Five variables gene regulatory network (Hlavacek and Savageau, 1996);

(b) Four variables didactic system (Voit and Almeida, 2004);

(c) Thirty variables system (Maki *et al.*, 2001);

(d) Cascade three variable system (Tsai and Wang, 2005);

(e) Yeast anaerobic fermentation pathway (Vera *et al.*, 2003);

(f) Three variable system (Voit, 2000a);

(g) Branched pathway with several feedback inhibition (Voit, 2000a);

(h) Anaerobic fermentation pathway in *Saccharomyces cerevisiae* (Curto et al., 1995);

(i) Kinetics model of ethanol fermentation (Wang *et al.*, 2001);

(j) Circadian oscillations of period protein in drosophila (Ingalls, 2004);

1.5 IDENTIFIABILITY ANALYSIS

As BST became a standard framework in representing biological systems such as metabolic networks, there is a strong interest in the development of methodologies to estimate unknown kinetic parameters from time-series experimental data efficiently. As shown in Table 1.1, many parameter estimation algorithms have been proposed that exploit the mathematical structure of standard models, like S-systems or generalized mass action (GMA) [69, 75-77]. Yet, many challenges still exist in this task as recently reviewed [60]. In this work, it is argued that the root of the problem faced in such parameter estimation problem arises from the lack of information in incomplete and noisy measurements in order to accurately estimate the model parameters. This is a parameter identifiability problem [78].

Identifiability analysis is basically concerned with the uniqueness of the parameters and is classified as *a priori* identifiability and practical identifiability. Before

actually proceeding to the experiment it is necessary to investigate whether, from the data that the experiment would generate (assuming the data to be ideal), it is theoretically possible to make unique estimates of all the unknown parameters. This is called *a priori* identifiability. This problem can also be viewed from experimental design point as follows: given the model structured is fixed, but that there is a choice of experiments, what sort of experiments should be designed in order to arrive at unique estimates for all the unknown parameters. Practical identifiability, a natural extension of *a priori* identifiability, deals with a measure of accuracy with which a parameter is estimated. To this end, a parameter identifiability analysis based on multivariate statistics is developed and applied this analysis to metabolic network models (see Chapter 2). Eventually, this analysis can be used to suggest model refinement and to optimize experimental design that maximizes the number of estimable parameters from data.

1.6 DESIGN OF EXPERIMENTS (DOE)

No two experiments will ever yield exactly the same data even when carried out under a strict protocol [79]. But one hopes that the model can explain the data within the experimental error, hence the model is valid. However, to validate a model an infinite number of experiments and data are required [80, 81]. Model invalidation is an integral part of model identification cycle, in which independent observations are used to test the closeness of model predictions and the data [82, 83]. At this stage of model identification cycle, the values of unknown parameters have been already been estimated and also a measure of accuracy for these estimates. In some situations these estimated parameters may be deemed adequate or identifiable for the intended purpose, in which case the

identification cycle can be terminated with model invalidation step. But in other cases, the model identification process has to be iterated again. A potential reason for this is that previous experiments have resulted in parameters whose accuracy is judged to be inadequate (unidentifiable) for the intended purpose. So in this step, design of experiments is carried out to yield a better experiment design which would generate a new set of data. One of the widely used designs is factorial design, which allows the simultaneous examination of the effects multiple independent variables and their degree of interaction. When the limitations of time and resources prevent the experimental exploration of all of the potentially feasible solutions for a certain process, the use of mathematical model could overcome this drawback. This is called as Model-Based Design of Experiments (MBDOE).

The first and foremost requirement for MBDOE is a preliminary model. Once a preliminary model is available, experimental degrees of freedom and constraints are defined. The main idea of MBDOE is to generate information-rich data that minimizes the parametric variances or maximizes the parameter precision. In this thesis, a curvature-based design criteria based on multi-objective optimization has been developed. The optimal experiment design obtained is combined with *in silico* experiments to obtain information-rich data. Based on these data, the model can be recalibrated and better parameter estimates can be obtained.

1.7 INTEGRATING IDENTIFIABILITY, PARAMETER ESTIMATION & DOE

The final step of this thesis was to integrate all the above mentioned tasks into a useful tool. Given the data and the model with nominal parameters, this tool will be able

to carry out the identifiability analysis, then perform optimal experiment design and finally get the better parameter estimates using the information-rich data. Although this concept is pretty well established, the novelty comes in the methodology of integrating the aforementioned tasks into an iterative process and the methods developed for identifiability analysis and MBDOE. The integration scheme is shown in Figure 1.2.

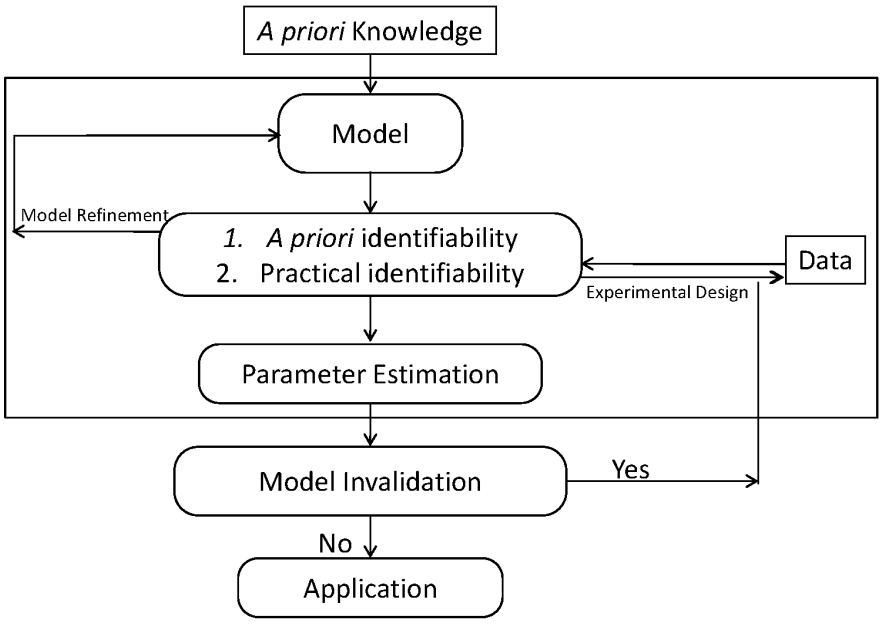

Figure 1.2. Iterative model building cycle to be performed by the proposed integrated tool

1.8 THESIS ORGANIZATION

This thesis concerns with the development of methods for identifiability analysis, design of experiments and iterative model identification of BST models. The main motivation of this thesis is to obtain a reliable model with the maximum parameter identifiability. Chapter 2 is concerned with the identifiability of such models, which in

this case is analyzed by the ability to uniquely or accurately identify parameter values from data. Two types of identifiability analyses: *a priori* and practical are developed in Chapter 2.

Chapter 3 deals with the application of identifiability analysis methods, developed in Chapter 2, to alternative BST formalisms. The first of half of this chapter focuses on the extension of the parameter identifiability methods to the decoupled models. The second part of this chapter deals with the identifiability analysis of linlog models.

Chapter 4 is dedicated to the model based design of experiments. This chapter commences with the general definition and purpose of design of experiments, followed by a literature survey of model-based design of experiments and the curvature-based DOE. In this chapter, a new multi-objective optimization based design criterion is proposed. The results and discussions follow where the new method is tested on a couple of examples taken from biotechnology industry**.**

Chapter 5 concerns with the integration of all the steps in the model identification cycle. This integration work is applied on a five-variable gene regulatory network modeled using the S-systems.

This thesis is concluded with Chapter 6 which presents an overview of significant findings and major contributions of this work. Future directions are also discussed.

CHAPTER 2†

IDENTIFIABILITY ANALYSIS OF BST MODELS

The extraction of parametric information from time series data is usually formulated as a minimization of the sum of squares of the differences between model simulation and data. The contributions from published works in this area usually differed in the formulation of the objective function, the optimization algorithms for finding the global minima, and/or the numerical methods for evaluating the objective function. Regardless of the objective function and the numerical algorithms used, a common problem faced in the parameter estimation of biochemical models is the existence of distinct parameter sets that give similar goodness of fit to experimental data. In essence, such problem is caused by the fact that (1) models are only an approximation of the true system and (2) data have limited information from which only a subset of parameters can be identified with sufficient accuracy. When fitting a mathematical model to experimental data, an important, but often overlooked issue, is the identifiability of parameters [84]. So, to obtain reliable results, parameter estimation should be complemented with identifiability analysis which assesses uniqueness of the estimated parameter values i.e. if other sets of values may be equally able to reproduce the available data. Such situations reduce the predictive abilities of the model. Identifiability of a dynamical model depends on the model structure, input-output functions, initial

† Excerpts of this work were published in
Srinath, S.; Gunawan, R., Parameter identifiability of power-law biochemical system models. *Journal of Biotechnology* **2010,** 149, (3), 132-140.

conditions [85] and the (unknown) true parameter values [86]. If parameter identifiability can be assessed before experimentation begins, then experiments can be designed to maximize the number of identifiable parameters.

Identifiability analysis is a well-established topic in statistics and systems engineering [78, 87]. The majority of models in systems biology are nonlinear and dynamic. So, checking the structural identifiability of this class of models is a daunting mathematical task. In recent years, the parameter identifiability problem has been a topic of major interest in systems biology [88-95]. Currently there are a few software packages for practical identifiability analysis: like PLA (Profile Likelihood Approach) [93] or AMIGO [96]. However, the software tool DAISY (Differential Algebra for Identifiability of SYstems [97] allows for structural identifiability but is limited by its size and functional form (only polynomial or rational) of the nonlinearities.

Model identifiability is defined as the ability to uniquely determine the model structure and parameters from a given set of experiments [98]. It is submitted that this definition lacks mathematical rigor, but as seen later, it has practical relevance in the development of identifiability analyses. Also, the "uniqueness" requirement can be relaxed when considering noisy measurements. In an ODE model,

$$\begin{aligned} \dot{\mathbf{X}}(t) &= \mathbf{f}(\mathbf{X}(t), \mathbf{u}(t), \boldsymbol{\theta}), \quad \mathbf{X}(0) = x_0 \\ \mathbf{y}(t) &= \mathbf{h}(\mathbf{X}(t), \mathbf{u}(t), \boldsymbol{\theta}). \end{aligned} \tag{2.1}$$

the model structure refers to the relationships or couplings among the states $\mathbf{X}$ as given in the RHS function $f(\mathbf{X}, \boldsymbol{\theta})$ [99]. In the above equation, $\mathbf{X} \in \mathbb{R}^{\mathbf{n}}, \mathbf{u} \in \mathbb{R}^{r}, \mathbf{y} \in \mathbb{R}^{p}$ and $\boldsymbol{\theta} \in \mathbb{R}^{q}$ are the states, inputs, outputs and parameter vector, respectively.

For example, in a reaction network, the model structure comprises the stoichiometric relationships (**N,** the stoichiometric matrix) and reaction flux equations (**v**), giving $\mathbf{f}(\mathbf{X},\boldsymbol{\theta}) = \mathbf{N}\mathbf{v}(\mathbf{X},\boldsymbol{\theta})$ [40]. The BST formalisms above are derived by using the power-law formalism for the reaction fluxes, such that the couplings among the metabolites or states (i.e. model structure) can now be inferred from the values of the parameters.

Following the definition above, parameter identifiability relates to whether the parameter values can be uniquely determined from data, and again this can be relaxed when considering noisy data. The topic of parameter identifiability is well established in mathematical modeling in science and engineering, including biotechnology [100-102]. Generally, there are two types of parameter identifiability; the first assumes perfect (noise-free) data, which is referred to as *structural* or *a priori* identifiability, while the second considers data quantity and quality, referred to as *practical* identifiability. The following sections present methods to study these identifiability conditions. This chapter is structured as follows. First, *a priori* and practical identifiability analyses will be discussed at length. In the subsequent section, the three methods of estimating confidence region and checking practical identifiability are presented with increasing simplifications.

2.1 *A PRIORI* IDENTIFIABILITY ANALYSIS

Identifiability, when perfect data is assumed has been a subject of research from 1970's and such an analysis was formally put forth by Bellman and Åström [99]. They referred to this kind of analysis as structural identifiability, but the term *a priori* identifiability analysis is also commonly used to address the same concept. The term "*a priori*" implies that the calculations can be done before a proposed experiment is carried

out, whereas the term "structural" can be taken to imply that the outcome should depend only on the model structure [100]. Although the terms *a priori* and structural identifiability analysis are often used interchangeably, there is a subtle difference in their definition.

2.1.1 Definition I (Structural Identifiability)

In the **limit of an infinite number of observations** [103] and **noise-free data** the model parameters are structurally identifiable if,

$$\begin{aligned} &\forall\ \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p,\ \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \Rightarrow \exists\ \mathrm{t} \text{ with} \\ &\mathbf{h}(\mathbf{x}(t, \boldsymbol{\theta}_1, u)) \neq \mathbf{h}(\mathbf{x}(t, \boldsymbol{\theta}_2, u)). \end{aligned} \tag{2.2}$$

The above definition for structural identifiability is often too strict. There can be situations where the parameters might not be identifiable under above mentioned conditions, but nevertheless would be identifiable for a reasonable restricted region of parameter space.

2.1.2 Definition II (*A priori* Identifiability)

The parameters $\boldsymbol{\theta}$ of a model are a priori or locally identifiable in a neighborhood of a parameter $\boldsymbol{\theta}_0$, if

$$\begin{aligned} &\forall\ \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \left\{ \boldsymbol{\theta} \in \mathbb{R}^p \;\middle|\; \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \varepsilon \right\},\ \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2 \\ &\Rightarrow \exists\ \mathrm{t} \text{ with } \mathbf{h}(\mathbf{x}(t, \boldsymbol{\theta}_1, u)) \neq \mathbf{h}(\mathbf{x}(t, \boldsymbol{\theta}_2, u)). \end{aligned} \tag{2.3}$$

The identifiability question in the presence of real noisy data, often referred to as practical or a posteriori identifiability, is classified and treated in the next section.

2.1.3 REVIEW OF EXISTING METHODS

Since Bellman and Åström [99] put forth the concept of identifiability, there has been many methods published for solving the structural or *a priori* identifiability problem. Some of these methods are discussed in this section. The details of the methods can also be found in [100, 104].

2.1.3.1 Taylor-series approach

In order to test the identifiability of the parameters of a model described by Eq. (2.1), it is required to be able to characterize its input-output behavior and thus to study the properties of its solutions for various inputs. In this approach, the output function **y** is expanded as a Taylor series and derivatives are evaluated at a specific time (usually t=0) in order to obtain mathematical expressions as simple as possible.

$$\mathbf{y}(t) = y(0) + t\frac{d\mathbf{y}}{dt}(0) + \frac{t^2}{2!}\frac{d^2\mathbf{y}}{dt^2}(0) + \ldots \tag{2.4}$$

The derivatives in Eq. (2.4) can be computed as functions of model parameters from Eq. (2.4)

$$\begin{aligned} y(0) &= \gamma_0(\theta) \\ \frac{dy}{dt}(0) &= \gamma_1(\theta) \\ &\ldots \end{aligned} \tag{2.5}$$

If one assumed that all the derivatives in Eq. (2.5) are known, then the subsequent step involves solving the model parameters as functions of the derivatives and inputs:

$$\begin{aligned}
\theta_1 &= \beta_1(\mathbf{y}(0), \frac{dy}{dt}(0), ..., u) \\
\theta_2 &= \beta_2(\mathbf{y}(0), \frac{dy}{dt}(0), ..., u) \\
&... \\
\theta_q &= \beta_q(\mathbf{y}(0), \frac{dy}{dt}(0), ..., u).
\end{aligned} \tag{2.6}$$

Analyzing the existence of the solution to Eq. (2.6) will give conclusions about the structural identifiability. If a single solution exists, then the model is deemed theoretically globally identifiable. But if more than one solution exists, then the model is only locally identifiable. Finally, if there are an infinite number of solutions possible, then the model is structurally unidentifiable. Although this method is simple it has the following drawbacks.

a) The number of derivatives to be computed increases with the number of parameters in the model, maintaining at least a 1:1 ratio. There is no limit for the number of derivatives to be computed [86].
b) If Eq. (2.6) can be solved, working out every parameter value can be very labor-intensive, even with the aid of dedicated software for symbolic computation [105, 106] such as MATLAB (Mathworks, Inc.), Maple (Maplesoft, Inc.) or Mathematica (Wolfram Research, Inc.)

2.1.3.2 Generating series approach

This next approach uses the relationship between the Lie derivatives and non-linear observability. This method requires the model to be linear in inputs $u(t)$ and allows the extension to the entire class of bounded and measurable input functions.

$$\frac{d\mathbf{x}}{dt} = f_0(\mathbf{x},\boldsymbol{\theta}) + \sum_{i=1}^{m} u_i(t)\, f_i(\mathbf{x},\boldsymbol{\theta}),$$
$$\mathbf{x}(0) = x_0(\boldsymbol{\theta})\; y(t,\boldsymbol{\theta}) = \mathbf{h}(\mathbf{x},\boldsymbol{\theta}). \tag{2.7}$$

This method is based on the model output function $\mathbf{h}(x,\theta)$ and their successive Lie derivatives, $L_{f_{j0}} \ldots L_{f_{jk}} \mathbf{h}(x,\boldsymbol{\theta})$, as evaluated at a specific time, where simple enough mathematical expressions can be obtained. The Lie derivative along a vector field f_i is equivalent to

$$L_{f_i} = \sum_{j=1}^{n} f_{j,i}(\mathbf{x},\boldsymbol{\theta}) \frac{\partial}{\partial x_j}. \tag{2.8}$$

where $f_{j,i}$ is the j-th component of f_i. For a given model, as in Eq. (2.7), a series expansion can be written based on Lie derivatives so that the model output can be expanded in series with respect to inputs and time denoted by the subscript 0.

Again, the conclusions regarding the structural identifiability of the model can be drawn from the number of solutions obtained for the parameters [107]. The mathematical expressions obtained using this method are much more simpler than the Taylor series method [108]. Many researchers used this method to compute the structurally identifiable parameters [109, 110] and even computational tools have been developed based on the generating series approach to compute the structurally identifiable parameters [111]. In [111] the authors have applied this method to several biological models. This method has a drawback in the sense that there is no way to know a priori how many Lie derivatives to be computed in order to obtain a determined system of equations, i.e. having same number of unknowns and equations.

2.1.3.3 Similarity transformation approach

This method is based on the local state isomorphism theorem and assumes that the entire class of bounded and measureable functions is available [86]. Controllability and observability are two important conditions to be fulfilled before applying this method. Controllability ensures that the system outputs can be chosen to reach any points in the state-space starting from any initial condition. Observability assures that every initial condition and hence every state profile can be estimated uniquely from the input-output measurements. The bottleneck of this method is to compute input trajectories that make the parameters identifiable from the measurements. If two states $\bar{\mathrm{x}}$ and $\tilde{\mathrm{x}}$ corresponding to two different parameter sets $\bar{\theta}$ and $\tilde{\theta}$, respectively, are considered, then the corresponding models will have the same input/output behavior for any input $u(t)$ and $t > 0$ if and only if local state isomorphism $\lambda : \mathrm{V} \to \mathbb{R}^{\mathrm{n}}$, $\bar{\mathrm{x}} \to \tilde{\mathrm{x}} = \lambda(\mathrm{x})$ (where V is a neighborhood of $\tilde{\mathrm{x}}(0)$) exists which fulfills the following conditions:

$$
\begin{aligned}
&\operatorname{rank}\left(\frac{\partial \lambda}{\partial \mathrm{x}^{\mathrm{T}}}\right)\Bigg|_{\mathrm{x}=\bar{\mathrm{x}}} = n \\
&\lambda(\bar{\mathrm{x}}(0)) = \tilde{\mathrm{x}}(0) \\
&f(\lambda(\bar{\mathrm{x}}), \theta) = \frac{\partial \lambda}{\partial \mathrm{x}^{\mathrm{T}}}\Bigg|_{\mathrm{x}=\bar{\mathrm{x}}} . f\left(\lambda(\bar{\mathrm{x}}), \bar{\theta}\right) \\
&g(\lambda(\bar{\mathrm{x}}), \theta) = \frac{\partial \lambda}{\partial \mathrm{x}^{\mathrm{T}}}\Bigg|_{\mathrm{x}=\bar{\mathrm{x}}} . g\left(\lambda(\bar{\mathrm{x}}), \bar{\theta}\right) \\
&h(\lambda(\bar{\mathrm{x}}), \theta) = h\left(\bar{\mathrm{x}}, \bar{\theta}\right).
\end{aligned}
\tag{2.9}
$$

If all solutions of the system (2.9) for $\bar{\theta}$ and $\bar{\mathrm{x}}$ can be uniquely obtained, the model is structurally globally identifiable [86].

2.1.3.4 Differential Algebra

All methods considered so far usually produce a system of algebraic equations to be solved for the parameters. Differential algebra, in which differentiation is added to the classical axioms of algebra, makes it possible to use a similar approach to eliminate the state variables so as to get differential input-output relations, only involving known variables and their derivatives and the parameters to be estimated, from which identifiability can be studied. This method reformulates model equations as linear regressions regarding parameters by using traditional algebraic operations jointly with differentiation [112]. The model requires non-zero, differentiable inputs and is only applicable to polynomial or rational models. The application of this method is described in [112] and a software implementation of it in [113]. However, it has been shown that differential algebra methods not only converge on the solution very slowly, but also fail with fairly large complex models [114]. As mentioned earlier, a software package, DAISY [97], based on this differential algebra approach is available for identifiability analysis of nonlinear models. But there is a limitation regarding the size and functional form of the nonlinear model (either polynomial or rational) that can be implemented in this software.

2.1.3.5 Hybrid Methods

Apart from the above mentioned methods, there also exist a few hybrid methods which combine the differential algebra with either generating series approach or Taylor series approach [115]. There are a few re-parametrization approaches which transform

the original models into structurally identifiable ones [106, 116]. However, these methods are not systematic and the physical meaning of the associated parameters might be lost.

All the above mentioned methods are analytical methods and in general applying them to practical problems is not an easy task. The tediousness of the algebraic manipulations involved makes computer algebra attractive. Breams *et al* [117] explored an alternative route, mainly based on numerical computation. They used interval analysis and interval propagation and this procedure enables checking if the system has a unique solution and hence is globally identifiable. However, long computational times limit this process, which is only acceptable for models up to three dimensions. Walter *et al* [118] modified this interval analysis procedure to quicken the algorithm's convergence. The modified algorithm is more efficient and able to handle complicated problems. Another numerical method for structural identifiability, proposed by Sedoglavic [119], is based on probabilistic polynomial-time algorithm which computes the set of observable variable of a model and gives the number of non-observable variables which should be assumed to be known in order to obtain observable system.

2.1.3.6 Profile Likelihood Approach

Timmer and co-workers presented a method based on profile-likelihood which enables to detect the structural and practical identifiabilities. In this method, a parameter is deemed identifiable if the confidence interval of its estimate is finite. The main idea of this approach is to explore the parameter region of each parameter in the direction of the least increase in χ^2.

$$\chi^2(\theta) = \left(\mathbf{y}_{\exp} - \mathbf{y}\right)' \Sigma^{-1} \left(\mathbf{y}_{\exp} - \mathbf{y}\right) \tag{2.10}$$

Profile likelihood (PL) χ^2_{PL} was used for the task of identifiability. For each parameter it is calculated by

$$\chi^2_{\mathrm{PL}}(\theta_i) = \min_{\theta_{j\neq i}} \left[\chi^2(\theta)\right] \tag{2.11}$$

re-optimizing $\chi^2(\theta)$ with respect to all parameters $\theta_{j\neq i}$, for each value of parameter θ_i. According to this method if a parameter is structurally non-identifiable it follows the functional relation, defined in Eq. (2.1), $h(\theta) = 0$. Similarly, practical identifiability detects the direction in which the likelihood flattens out.

2.1.3.7 Proposed method

The method described below is based on the first order derivatives of the model (measured) output with respect to the parameters, also called the sensitivity matrix $\hat{\mathbf{F}}$. [120-122]. The sensitivity matrix reflects how much changes in the parameter values will affect the output. If the outputs have zero sensitivity with respect to a parameter, then intuition tells that this parameter cannot be estimated from the output. Similarly, if two parameters cause proportional changes to the output, i.e. the sensitivity to one parameter is a constant multiple of the other, then intuition also suggests that the data cannot differentiate the two parameters. To put this in mathematical terms, one can start from the inverse modeling problem; that is, given a generic model:

$$\mathbf{y} = F(\boldsymbol{\theta}; \mathbf{x}). \tag{2.12}$$

where **x** and **y** denote the measured input/state and output variables, respectively, then the inverse problem can be thought as finding an inverse of this model such that

$$\boldsymbol{\theta} = F^{-1}\left(\mathbf{y};\mathbf{x}\right). \tag{2.13}$$

If the model $F(\boldsymbol{\theta};\mathbf{x})$ is invertible, then the parameters $\boldsymbol{\theta}$ can be determined uniquely, i.e. the parameters are *a priori* identifiable. Now, by taking the Taylor series expansion of $F(\boldsymbol{\theta};\mathbf{x})$ around the nominal parameters $\boldsymbol{\theta}^*$, the *a priori* identifiability can be studied from the sensitivity matrix based on:

$$\begin{aligned} \mathbf{y} &= F\left(\boldsymbol{\theta}^*;\mathbf{x}\right) + \left.\frac{\partial y}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}^*;\mathbf{x}} \left(\boldsymbol{\theta} - \boldsymbol{\theta}^*\right) + \ldots \\ \boldsymbol{\theta} - \boldsymbol{\theta}^* &\approx \left.\frac{\partial y}{\partial \boldsymbol{\theta}}\right|^{-1}_{\boldsymbol{\theta}^*;\mathbf{x}} \left(\mathbf{y} - F\left(\boldsymbol{\theta}^*;\mathbf{x}\right)\right) = \hat{\mathbf{F}}_{\cdot}^{-1}\left(\boldsymbol{\theta}^*;\mathbf{x}\right)\left(\mathbf{y} - F\left(\boldsymbol{\theta}^*;\mathbf{x}\right)\right). \end{aligned} \tag{2.14}$$

Thus, the model invertibility is approximated by the invertibility of the sensitivity matrix $\hat{\mathbf{F}}_{\cdot}$. Notice that zero sensitivity or linearly dependent sensitivities will cause $\hat{\mathbf{F}}_{\cdot}$ to become singular and thus cannot be inverted, in agreement with the intuitions above. The sensitivity matrix is formed by stacking the sensitivity coefficients with respect to time. The sensitivity matrix is given by

$$\hat{\mathbf{F}}_{\bullet} = \begin{pmatrix} \left.\frac{\partial y_1}{\partial \theta_1}\right|_{t=t_1} & \cdots & \left.\frac{\partial y_1}{\partial \theta_P}\right|_{t=t_1} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial y_R}{\partial \theta_1}\right|_{t=t_1} & \cdots & \left.\frac{\partial y_R}{\partial \theta_P}\right|_{t=t_1} \\ \left.\frac{\partial y_1}{\partial \theta_1}\right|_{t=t_2} & \cdots & \left.\frac{\partial y_1}{\partial \theta_P}\right|_{t=t_2} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial y_R}{\partial \theta_1}\right|_{t=t_2} & \cdots & \left.\frac{\partial y_R}{\partial \theta_P}\right|_{t=t_2} \end{pmatrix}. \quad (2.15)$$

Following this argument, the *a priori* identifiable parameters are determined using an algorithm that ranks linearly independent columns of the sensitivity matrix according to the Euclidean norms. The pseudo-algorithm is summarized as follows: [84]

1. identify the column of $\hat{\mathbf{F}}_{\bullet}$ with the highest magnitude as measured by the Euclidean norm,
2. if the magnitude exceeds a certain threshold, remove this column from $\hat{\mathbf{F}}_{\bullet}$ and assign the corresponding parameter as *a priori* identifiable; if the magnitude is smaller than the threshold, then assign this and all remaining parameters as **not** *a priori* identifiable.
3. find the projection of this column vector on the space of the remaining $\hat{\mathbf{F}}_{\bullet}$ and subtract this projection from $\hat{\mathbf{F}}_{\bullet}$,
4. repeat from 1.

Another approach that could be done involves calculating the correlation matrix based on $\hat{\mathbf{F}}_{\bullet}$, which has been applied to *in silico* gene network [4, 123].

2.2 PRACTICAL IDENTIFIABILITY ANALYSIS

In practice, experimental data are not perfect, and despite the availability of high throughput technology, biological data are known to be noisy [124]. Such noise contaminates the true signal and reduces the degree of information in the data. The estimation procedure can provide measures of the goodness of fit between model response and experimental data as well as a measure of accuracy of all estimated parameters. A measure of the accuracy with which the unknown parameters are estimated is of particular importance. This is referred to as practical or *a posteriori* identifiability analysis and can be considered as a natural extension to *a priori* identifiability analysis. But, even *a priori* identifiable parameters cannot be determined to an infinite accuracy, i.e. an infinite number of significant figures. So, the uniqueness criterion in the definition of identifiability needs to be relaxed and *practical* identifiability will be judged based on the uncertainty in the parameter estimates, which arises due to the aforementioned data noise.

2.2.1 PROPOSED METHODS

A parameter estimation problem is typically formulated as a weighted least square minimization problem, given by:

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \left(\mathbf{y}_{obs} - F(\boldsymbol{\theta};\mathbf{x})\right)^T \mathbf{W}\left(\mathbf{y}_{obs} - F(\boldsymbol{\theta};\mathbf{x})\right) \\ &\text{subject to} \\ &G(\boldsymbol{\theta}) \le 0 \\ &H(\boldsymbol{\theta}) = 0. \end{aligned} \tag{2.16}$$

where the data $\mathbf{y}_{obs} = F(\boldsymbol{\theta};\mathbf{x}) + \boldsymbol{\varepsilon}$ are assumed to be contaminated with independent and identically distributed (i.i.d.), zero mean and constant variance noise $\boldsymbol{\varepsilon}$, $F(\boldsymbol{\theta};\mathbf{x})$ is the model prediction, $\mathbf{W}$ denotes the weighting matrix (refer to Appendix A for more details). For numerical and physical reasons, the parameter search is often constrained to a limited, feasible region in the parameter space, which is defined by inequality and equality constraints on $G(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$, respectively. If the inverse of noise covariance matrix $\mathbf{V}$ is selected as the weights, $\mathbf{W} = \mathbf{V}^{-1}$, then the problem above is equivalent to the maximum likelihood estimation under Gaussian noise assumption. Since the noise $\boldsymbol{\varepsilon}$ is random, the parameter estimates are also random, and the degree of parametric uncertainty can be estimated as a function of noise variance. This topic is well-studied in the subject of nonlinear regression [87]. Specifically, the practical identifiability here is formulated as a test on the confidence interval or region of the parameter estimates. The confidence region is defined as a subset of the feasible parameter space with a specified probability, given by the confidence level, of containing the true parameter values. Note that the confidence region is defined around the (random) parameter estimates and thus this region is random.

In this work, a parameter θ is deemed practically identifiable when the parametric confidence region does not cross the axis $\theta = 0$. The confidence level is often taken to be 95%, which is approximately the probability between mean ± 2 times standard deviation for a Gaussian random variable. The zero-axis crossing test is commonly used to evaluate practical identifiability in other studies [123]. In the context of BST, the value $\theta = 0$ has a physical significance. First, the rate constants are assumed to be positive, i.e. the flux is irreversible, and a reversible reaction is usually expressed as two irreversible reactions.

Second, the sign of the kinetic order parameters also has a physical interpretation, in which a positive value indicates substrate or activation and a negative implies inhibition [44]. Hence, if the parametric uncertainty crosses the $\theta = 0$ axis, then either the directionality of the flux or the influence of one metabolite on another cannot be determined from the data to the specified confidence level. Note that when the true parameter value is zero, the criterion cannot adequately address the practical identifiability of this parameter. Instead a hypothesis test with the null hypothesis $\theta = 0$ for the suspect parameters can be done, which upon a failure to reject the null hypothesis, gives a (weak) support that the true value is practically identifiable as zero.

In what follows, three methods of estimating confidence region and checking practical identifiability are presented with increasing simplifications. Before going into the mathematical derivations, consider a less rigorous but a more practical definition of the parameter confidence region. Given the minimum solution of the parameter estimation problem in Eq.(2.16), the parameter values in the neighborhood of the minima $\hat{\boldsymbol{\theta}}$ should also give a good approximation of the data with a low $\Phi(\boldsymbol{\theta})$. Since data noise is random, the true parameter values $\boldsymbol{\theta}^*$ may not even give the lowest $\Phi(\boldsymbol{\theta})$, but are still expected to lie near the optima. Thus, it would seem reasonable to define the confidence region of the parameter based on the contours of $\Phi(\boldsymbol{\theta})$, for example: [125]

$$\left\{\boldsymbol{\theta} : \Phi(\boldsymbol{\theta}) < c\Phi\left(\hat{\boldsymbol{\theta}}\right)\right\}. \tag{2.17}$$

for a constant $c > 1$. Although the exact value of c for a general nonlinear parameter estimation is unknown, its approximate value based on simplifying assumption can be

derived. In the development of the methods below, the noise is assumed to be i.i.d. random sample from a Gaussian distribution with zero mean and the covariance matrix **V**. Alternatively, one can also use a Taylor series expansion of $\Phi(\boldsymbol{\theta})$, in which the Hessian can give an approximation of the $\Phi(\boldsymbol{\theta})$ contour, as recently done [126].

2.2.1.1 Method 1

The first method uses the following asymptotic approximation of 100(1-α)% confidence region: [125]

$$\left\{\boldsymbol{\theta}:\Phi(\boldsymbol{\theta})<\Phi(\hat{\boldsymbol{\theta}})\left(1+\frac{p}{n-p}F^{\alpha}_{p,n-p}\right)\right\}. \tag{2.18}$$

where n and p denote the number of data points and parameters, respectively and $F^{\alpha}_{p,n-p}$ is the percentage point F-distribution with p and $n–p$ degrees of freedoms. The zero crossing criterion is formulated as a constrained minimization problem given by:

$$\begin{aligned}&\min_{\boldsymbol{\theta}}\operatorname{sign}(\hat{\theta}_i)\theta_i\\&\text{subject to}\\&\Phi(\boldsymbol{\theta})<\Phi(\hat{\boldsymbol{\theta}})\left(1+\frac{p}{n-p}F^{\alpha}_{p,n-p}\right)\\&g(\boldsymbol{\theta})\le 0\\&h(\boldsymbol{\theta})=0.\end{aligned} \tag{2.19}$$

and the i-th parameter θ_i is said to be practically identifiable when the minima is positive. Note that the identifiability check is done for parameter space bounded by the confidence region defined within the feasible parameter space.

2.2.1.2 Method 2

The second method applies a linear approximation of the model using a Taylor series expansion as in Eq. (2.14). Substituting this to Eq. (2.18) and rearranging the terms, the (linearized) 100(1-α)% confidence region is a hyper ellipsoidal parameter space defined by:

$$\left\{\boldsymbol{\theta}:\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^{T}\left(\hat{\mathbf{F}}_{\cdot}^{T}\mathbf{WS}\right)\left(\hat{\mathbf{F}}_{\cdot}^{T}\mathbf{WVW}\hat{\mathbf{F}}_{\cdot}\right)^{-1}\left(\hat{\mathbf{F}}_{\cdot}^{T}\mathbf{W}\hat{\mathbf{F}}_{\cdot}\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)<\frac{p}{n-p}\Phi\left(\hat{\boldsymbol{\theta}}\right)F_{p,n-p}^{\alpha}\right\}. \tag{2.20}$$

Note that the covariance of the parameter estimates for the linearized model is given by $\mathbf{V}_{\theta}$ as follows: [127]

$$\mathbf{V}_{\theta}=\left(\hat{\mathbf{F}}_{\cdot}^{T}\mathbf{W}\hat{\mathbf{F}}_{\cdot}\right)^{-1}\hat{\mathbf{F}}_{\cdot}^{T}\mathbf{WVW}\hat{\mathbf{F}}_{\cdot}\left(\hat{\mathbf{F}}_{\cdot}^{T}\mathbf{W}\hat{\mathbf{F}}_{\cdot}\right)^{-1}. \tag{2.21}$$

For the maximum likelihood estimation, i.e. $\mathbf{W} = \mathbf{V}^{-1}$, the confidence region further reduces to

$$\left\{\boldsymbol{\theta}:\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^{T}\left(\hat{\mathbf{F}}_{\cdot}^{T}\mathbf{V}^{-1}\hat{\mathbf{F}}_{\cdot}\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)<\frac{p}{n-p}\Phi\left(\hat{\boldsymbol{\theta}}\right)F_{p,n-p}^{\alpha}\right\}. \tag{2.22}$$

where the term $\hat{\mathbf{F}}_{\cdot}^{T}\mathbf{V}^{-1}\hat{\mathbf{F}}_{\cdot}$ is also known as the Fisher information matrix (FIM) [128, 129]. In this method the existence of an intersection between the hyper plane $\theta_i = 0$ and confidence region (CR) given in Eq. (2.22) is checked. Without loss of generality, the derivation of the CR identifiability test for the first parameter θ_1 is given below. Recall that θ_1 is not practically identifiable if there exists a solution to Eq. (2.22). Hence, the identifiability analysis can be written as:

$$\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^T \mathbf{V}_\theta^{-1}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right) < ps^2 F_{p,n-p}^{\alpha}$$

$$\Rightarrow \begin{bmatrix} -\theta_1 \\ \boldsymbol{\theta}_{\setminus 1}-\hat{\boldsymbol{\theta}}_{\setminus 1} \end{bmatrix}^T \begin{bmatrix} \lambda_{\setminus 1} & \boldsymbol{\lambda}_{\setminus 1}^T \\ \boldsymbol{\lambda}_{\setminus 1} & \boldsymbol{\Lambda}_{\setminus 1} \end{bmatrix} \begin{bmatrix} -\theta_1 \\ \boldsymbol{\theta}_{\setminus 1}-\hat{\boldsymbol{\theta}}_{\setminus 1} \end{bmatrix} < ps^2 F_{p,n-p}^{\alpha}$$

$$\Rightarrow \lambda_{\setminus 1}\hat{\theta}_1^2 - \hat{\theta}_1\boldsymbol{\lambda}_{\setminus 1}^T\left(\boldsymbol{\theta}_{\setminus 1}-\hat{\boldsymbol{\theta}}_{\setminus 1}\right) - \left(\boldsymbol{\theta}_{\setminus 1}-\hat{\boldsymbol{\theta}}_{\setminus 1}\right)^T \boldsymbol{\lambda}_{\setminus 1}\hat{\theta}_1 + \left(\boldsymbol{\theta}_{\setminus 1}-\hat{\boldsymbol{\theta}}_{\setminus 1}\right)^T \boldsymbol{\Lambda}_{\setminus 1}\left(\boldsymbol{\theta}_{\setminus 1}-\hat{\boldsymbol{\theta}}_{\setminus 1}\right) < ps^2 F_{p,n-p}^{\alpha} \qquad (2.23)$$

$$\Rightarrow \lambda_{\setminus 1}\hat{\theta}_1^2 + 2\boldsymbol{\lambda}_{\setminus 1}^T\hat{\boldsymbol{\theta}}_{\setminus 1}\hat{\theta}_1 + \hat{\boldsymbol{\theta}}_{\setminus 1}^T\boldsymbol{\Lambda}_{\setminus 1}\hat{\boldsymbol{\theta}}_{\setminus 1} - 2\hat{\theta}_1\boldsymbol{\lambda}_{\setminus 1}^T\boldsymbol{\theta}_{\setminus 1} - 2\hat{\boldsymbol{\theta}}_{\setminus 1}\boldsymbol{\Lambda}_{\setminus 1}\boldsymbol{\theta}_{\setminus 1} + \hat{\boldsymbol{\theta}}_{\setminus 1}\boldsymbol{\Lambda}_{\setminus 1}\boldsymbol{\theta}_{\setminus 1} < ps^2 F_{p,n-p}^{\alpha}$$

$$\Rightarrow \Psi + 2\hat{\theta}_1\boldsymbol{\lambda}_{\setminus 1}^T\boldsymbol{\theta}_{\setminus 1} + 2\hat{\boldsymbol{\theta}}_{\setminus 1}\boldsymbol{\Lambda}_{\setminus 1}\boldsymbol{\theta}_{\setminus 1} - \hat{\boldsymbol{\theta}}_{\setminus 1}\boldsymbol{\Lambda}_{\setminus 1}\boldsymbol{\theta}_{\setminus 1} > 0.$$

where

$$\Psi = \lambda_{\setminus 1}\hat{\theta}_1^2 + 2\boldsymbol{\lambda}_{\setminus 1}^T\hat{\boldsymbol{\theta}}_{\setminus 1}\hat{\theta}_1 + \hat{\boldsymbol{\theta}}_{\setminus 1}^T\boldsymbol{\Lambda}_{\setminus 1}\hat{\boldsymbol{\theta}}_{\setminus 1} - ps^2 F_{p,n-p}^{\alpha}$$

The Schur complement lemma states that the two statements below are equivalent: [130]

(1) $\mathbf{R}(x) > 0$ & $\mathbf{Q}(x) - \mathbf{S}(x)\mathbf{R}^{-1}(x)\mathbf{S}(x)^T > 0.$

$(2)\begin{pmatrix} \mathbf{Q}(x) & \mathbf{S}(x) \\ \mathbf{S}(x)^T & \mathbf{R}(x) \end{pmatrix} > 0$.

Under the assumption that noise is i.i.d., the analysis is performed only for the *a priori* identifiable parameters (AIP) such that FIM is full rank and its inverse is positive definite. Applying the Schur complement to Eq.(2.23), the practical identifiability of θ_1 is equivalent to the feasibility problem of the following LMI:

$$\begin{pmatrix} -\Psi + 2\hat{\theta}_1\boldsymbol{\lambda}_{\setminus 1}^T\boldsymbol{\theta}_{\setminus 1} + 2\hat{\boldsymbol{\theta}}_1^T\boldsymbol{\Lambda}_{\setminus 1}\boldsymbol{\theta}_{\setminus 1} & \boldsymbol{\theta}_{\setminus 1}^T \\ \boldsymbol{\theta}_{\setminus 1} & \boldsymbol{\Lambda}_{\setminus 1}^{-1} \end{pmatrix} > 0. \qquad (2.24)$$

The feasibility problem can be solved using off-the-shelf software, for example using the Robust Control toolbox in MATLAB. So, if the LMI is feasible, then it means that there exist at least one θ_i which satisfies the Eq. (2.22) and the hence the corresponding parameter is unidentifiable.

2.2.1.3 Method 3

While the above two methods consider multivariate statistical inference on the parameters, the confidence region can be further simplified into a single parameter axis, which is aptly called confidence interval. Again using linearization of the model, the (two-sided) 100(1-α)% confidence interval of the parameter θ_i is given by: [131]

$$\theta_i \pm t_{\alpha/2,n-p}\sqrt{\mathbf{V}_\theta(i,i)}. \tag{2.25}$$

where $\mathbf{V}_\theta$(i,i) is the i-th diagonal element of the parameter covariance matrix and $t_{\alpha/2,n-p}$ is the percentage point of the t-distribution.

The methods above describe three different approximations of the parameter confidence regions, as illustrated in Figure 2.1. The first method assumes that the residuals $\mathbf{e} = \mathbf{y}_{obs} - F(\boldsymbol{\theta};\mathbf{x})$ are asymptotically Gaussian (as $n \to \infty$), leading to the use of the F-distribution. As shown in Figure 2.1, the parametric nonlinearity is still captured by the first confidence region, while using linearized model, the second region is depicted as an ellipsoid. Finally, the third method produces a box confidence region, which often underestimates the true parametric uncertainty.

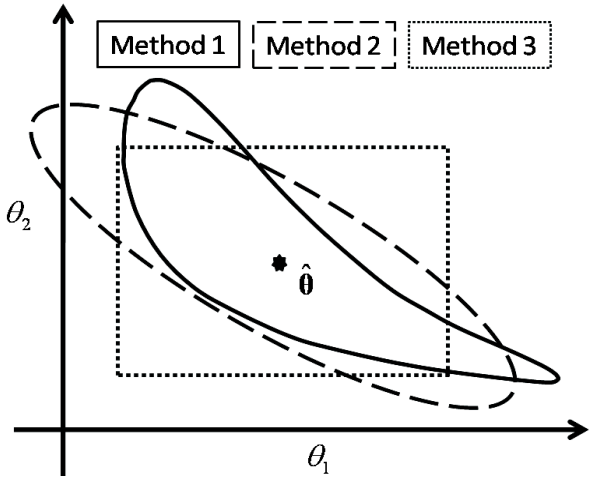

Figure 2.1. Comparison of Confidence Regions for Practical Identifiability

2.3 RESULTS AND DISCUSSION

In this section, the identifiability analyses explained above are applied to two inverse modeling problems: a GMA model of *L. lactis* [60] and an S-system model of *E. coli* metabolism [132]. The quantity and quality of the data in these models arguably represent the best-case scenario of an inverse modeling problem in this area, where time-series concentrations of all metabolites within the subsystem are available and the model dimensionality is small (<10). The analysis calculations were performed in MATLAB.

2.3.1 CASE STUDY I: GLYCOLYTIC PATHWAY IN *L. LACTIS*

The first case study deals with the glycolytic pathway which is depicted in Figure 2.2. The corresponding GMA model is given below. This pathway which converts the sugars into pyruvate is also called Embden-Meyerhof-Parnas pathway. The lactic acid bacteria or *L. lactis* has extensively been used in the production of buttermilk, cheese and yogurt. The bacteria are well characterized and the genome has been sequenced [133] and

thus making it a preferred choice of model for further research, like in [134]. This pathway is modeled using the GMA framework, within BST. The details of the model used and the experimental setup involved in obtaining the data are given in [60]. The regulation of glycolysis in *L. lactis* has been a subject of research since 1980's. The key enzymes in this pathway, phoshpfructokinase, fructose 1,6-biphosphate (FBP) aldolase, glyceraldehydes 3-phosphate dehydrogenase (GAPDH), pyruvate kinase (PK) and lactate dehydrogenase (LDH) were characterized and concentrations of several intermediates of this pathway had been already obtained using nuclear magnetic resonance (NMR). Neves *et al* [135] used this method to monitor the pools of labeled metabolites and end products, with a time resolution of 30 seconds, in a non-growing *L. lactis* cell suspensions following a bolus of ^{13}C-labelled glucose. In vivo experiments were performed using a circulating system described in [136]. [6-^{13}C] glucose (20mM) was supplied to the cell suspension and the time courses for substrate consumption, product formation and intracellular metabolite pools were monitored. Further details on extraction and quantification of the end products are available in [135].

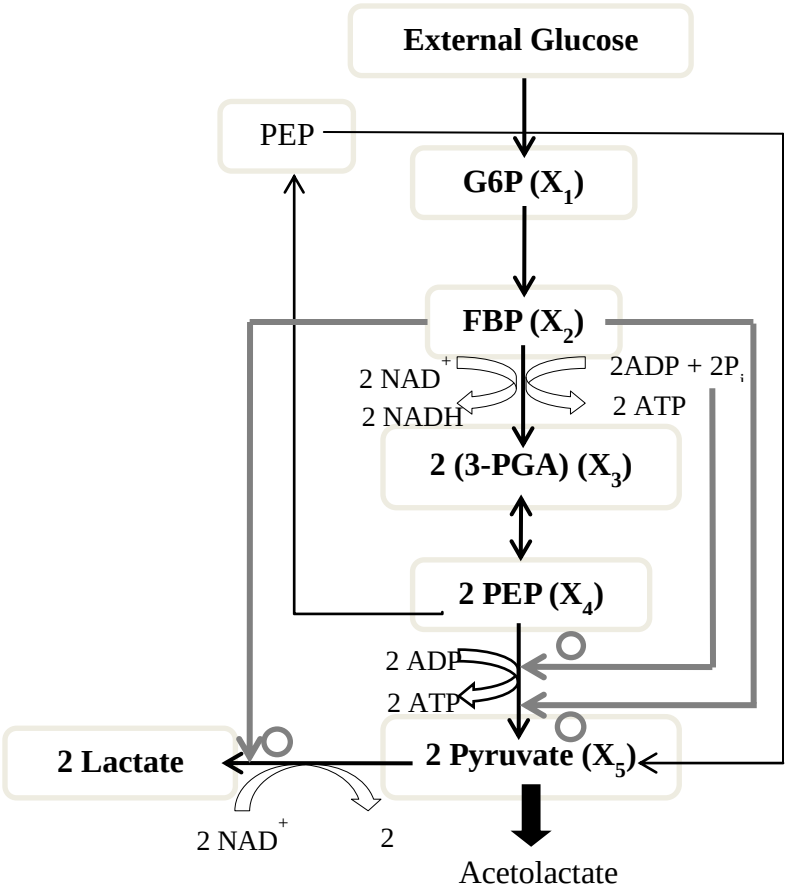

Figure 2.2. Glycolytic pathway in *L. lactis*.

A GMA model was proposed and fit to the data [60]. A combined bottom-up and top-bottom approach was used to infer the network topology. The model consists of 6 ODEs and 3 input variables. The inverse modeling problem consisted of 25 unknown parameters with 9 rate constants and 16 kinetic orders using NMR measurements of the metabolites, glucose substrate and other co-factors from a glucose perturbation experiment [60]. The model equations are given below.

$$
\begin{aligned}
\frac{dX_1}{dt} &= \alpha_1 Glc^{g_{1,Glc}} X_1^{g_{11}} X_4^{g_{14}} - \beta_1 X_1^{h_{11}} ATP^{h_{1,ATP}} \\
\frac{dX_2}{dt} &= \beta_1 X_1^{h_{11}} ATP^{h_{1,ATP}} - \beta_2 X_2^{h_{22}} P_i^{h_{2,Pi}} \\
\frac{dX_3}{dt} &= 2\beta_2 X_2^{h_{22}} P_i^{h_{2,Pi}} + \alpha_3 X_4^{g_{34}} - \beta_3 X_3^{h_{33}} \\
\frac{dX_4}{dt} &= \beta_3 X_3^{h_{33}} - \alpha_1 Glc^{g_{1,Glc}} X_1^{g_{11}} X_4^{g_{14}} - \alpha_3 X_4^{g_{34}} - \beta_{41} X_2^{h_{412}} X_4^{h_{414}} P_i^{h_{41,Pi}} - \beta_{42} X_4^{h_{424}} \\
\frac{dX_5}{dt} &= \alpha_1 Glc^{g_{1,Glc}} X_1^{g_{11}} X_4^{g_{14}} + \beta_{41} X_2^{h_{412}} X_4^{h_{414}} P_i^{h_{41,Pi}} - \beta_{51} X_5^{h_{515}} X_2^{h_{512}} - \beta_{52} X_5^{h_{525}} \\
\frac{dX_6}{dt} &= \beta_{51} X_5^{h_{515}} X_2^{h_{512}}.
\end{aligned}
\tag{2.26}
$$

The parameter estimation of this model has been attempted using a combination of least square and slope-based estimation [126] and linlog estimation [56] with various degrees of success. However, the parameter estimates from different methods were not in agreement, and as noted before, manual fitting can also give a reasonable approximation of the data [60]. The existence of multiple parameter sets giving similar data fit can be taken as a sign of identifiability issues, motivating the application of the above mentioned analyses

The study of *a priori* identifiability used the least square parameter estimates (**W = I,** refer Appendix A for other forms of least squares) as reported previously [60]. Figure 2.3 gives the total *a priori* identifiable parameters (AIP) as a function of the data sampling rate and experimental conditions. As expected, the number of AIP increased with the sampling rate and the number of glucose perturbations, since the degree of information in the dataset correspondingly increased. However, there was a diminishing return of AIP, since the additional dynamical information from more data points and additional perturbations is marginal at high sampling rate. At the maximum information,

only 19 out of 25 parameters were *a priori* identifiable, among which four are rate constants and 15 are kinetic orders. Using the sampling rate of 60 h^{-1} and a single glucose bolus as done in the original publication [60], the total number of AIP lowered to 15 (4 rate constants and 11 kinetic orders).

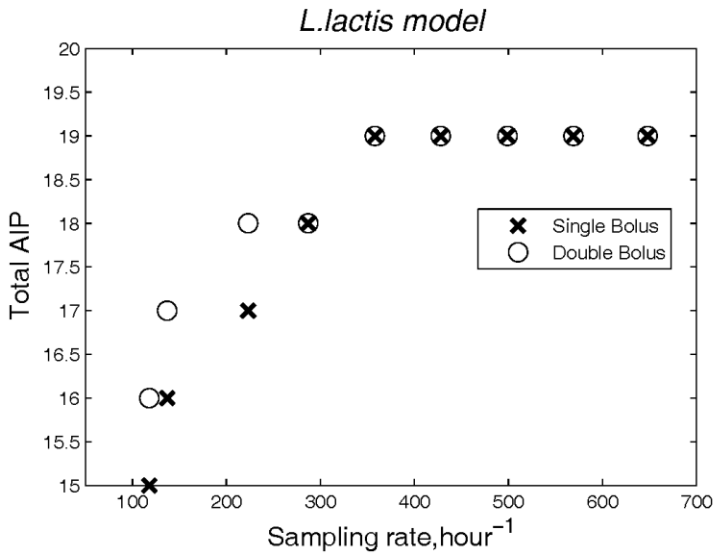

Figure 2.3. Effect of Sampling Rate and Experiments on *a priori* Identifiability for *L. lactis* model.

The practical identifiability analyses were done based on a 95% confidence level. Because of the difficulty in evaluating the objective function $\Phi(\boldsymbol{\theta})$ for some combinations of parameter values in the neighborhood of $\hat{\boldsymbol{\theta}}$ (stiffness problem [60]), only linearized method 2 and 3 of practical identifiability were applied to the AIP. The dynamical sensitivity matrix **S** was then computed only for the AIP (15 out of 25) using the direct differential method [122]. Finally, the noise variance in each metabolite measurement was estimated from the residuals (see further discussion below). In this case, the ellipsoidal confidence regions crossed the $\theta_i = 0$ axis for 10 out of the 15 AIPs, i.e. 5 of the AIPs are practically identifiable, all of which were kinetic order parameters. The

confidence interval analysis also concluded that the same 5 out of 15 AIPs were practically identifiable. Table 2.1 summarizes the identifiability results presented above for the *L. lactis* GMA model (see Table A in the Appendix A for detailed results).

Table 2.1: Summary of identifiability results for both the models

L. lactis* Model					
	Total Parameters	**AIP**	**Practical Identifiability**		
			Method 1	**Method 2**	**Method 3**
Rate constants	9	4	–	0	0
Kinetic order	16	11	–	4	5
***E. coli* Model#**					
Rate constants	10	5	5	3	5
Kinetic order	21	10	1	7	6

* At a sampling rate of **60 *per hour***, a single bolus of Glucose

\# At a sampling rate of **1 *per hour***, 40 and 50 g/L Glucose data

2.3.2 CASE STUDY 2: MODELING RECOMBINANT *E. COLI* GROWTH

The second case study describes the growth of a recombinant *E. coli* strain BL21/pBAW2 under two initial glucose concentrations (40 g/L and 50 g/L) [132]. The S-system model consists of 5 state variables, all of which were measured, on a hourly basis up to 14 hours, until glucose was fully consumed. The experiments were repeated to give two replicates of data for each initial glucose concentration. Three S-system models were proposed, of which the best (model III in [132]) is used in the analysis below. The

estimation of 31 model parameters was formulated as a weighted least square with time and maximum concentration values as the weights [132].

$$\begin{aligned}
\frac{dX_1}{dt} &= \alpha_1 X_1^{g_{11}} X_2^{g_{12}} - \beta_1 X_1^{h_{11}} X_2^{h_{12}} \\
\frac{dX_2}{dt} &= \alpha_2 - \beta_2 X_1^{h_{21}} X_2^{h_{22}} \\
\frac{dX_3}{dt} &= \alpha_3 X_1^{g_{31}} X_2^{g_{32}} - \beta_3 X_1^{h_{31}} X_2^{h_{32}} X_3^{h_{33}} \\
\frac{dX_4}{dt} &= \alpha_4 X_1^{g_{41}} X_2^{g_{42}} - \beta_4 X_1^{h_{41}} X_2^{h_{42}} X_4^{h_{44}} \\
\frac{dX_5}{dt} &= \alpha_5 X_1^{g_{51}} X_2^{g_{52}} - \beta_4 X_1^{h_{51}} X_2^{h_{52}} X_5^{h_{55}}.
\end{aligned} \tag{2.27}$$

where, X_1 is the concentration of cell mass, X_2 is the concentration of glucose, X_3 is the concentration of protein, X_4 is the concentration of lactate and X_5 is the concentration of acetate. Thus, the cell mass and energy are included into the dynamical equations for protein, lactate and acetate.

Figure 2.4 shows the number of AIP as a function of sampling rate and the experimental conditions. As expected, the number of AIP increases with increasing data informativeness brought by increasing sampling rate and additional experiments, again with diminishing returns. Using data from two experimental conditions, the maximum AIP was 24 out of 31, of which 6 are rate constants and 18 are kinetic orders. For the actual experimental settings performed, only 15 parameters were found to be *a priori* identifiable, comprising 5 rate constants and 10 kinetic orders.

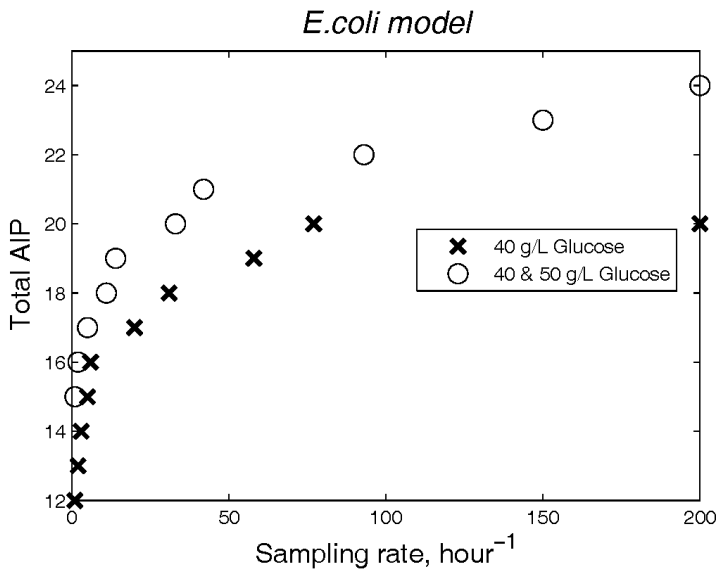

Figure 2.4. Effect of Sampling Rate and Experiments on *a priori* Identifiability for *E. coli* model.

The application of linearized practical identifiability analyses using a 95% confidence level for the AIP indicated that 11 out of 15 AIP (5 rate constants and 6 kinetic orders) were practically identifiable based on the confidence interval, while this number dropped to 7 (2 rate constants and 5 kinetic orders) based on the ellipsoidal confidence region. Following the modified collocation method used in the original publication [132], the objective function evaluation $\Phi(\theta)$ now becomes algebraic, allowing the use of method 1 above. Accounting for parametric nonlinearity in the confidence region, method 1 gave an even lower number of practically identifiable parameters (PIP) of only 6 (5 rate constants and 1 kinetic order). These results are summarized in Table 2.1.

Note that method 1 was applied to all parameters, not just the AIP like in the applications of method 2 and 3. The PIP from method 1 was found to be among the AIP as expected (see Table B in the Appendix A). The differences between the nonlinear and

linearized practical identifiability tests demonstrate the effect of parametric nonlinearity. As rate constants show up linearly in the model, linearized analysis like method 2 and 3 can sufficiently describe the parametric uncertainty. In this case, however, method 2 was likely to overestimate the parametric confidence region, giving a lower number of practically identifiable rate constants. On the other hand, kinetic order parameters appear in the exponents, and thus methods based on linearized model may only be accurate in a small neighborhood of the parameter estimates.

2.3.3 DISCUSSION

The results of the parametric identifiability study above explain that the fundamental challenge faced in the inverse modeling of the BST is rooted from the lack of complete parameter identifiability. This implies that the simultaneous estimation of all parameters from the experimental data is an ill-conditioned problem. Even when the pertinent metabolites, substrates and co-factors are measured, not all (~80%) of the parameters are identifiable from noise-free (perfect) data, and only 50-60% of the parameters are *a priori* identifiable from the typical experimental settings (sampling rate between 1 min^{-1} to 1 h^{-1}). The AIP forms a superset of the identifiable parameters in practice when data are noisy, and thus the number of AIP gives the upper bound for the number of PIP. Using different approximations of the parametric confidence regions, the fraction of AIP that is practically identifiable is estimated to be about 50%, while the actual number of PIP could be lower than the current estimates. Thus, the overall percentage of the parameters that are estimable from noisy experimental data can potentially be lower than 25%.

Despite the bleak picture painted by this study, the inverse modeling is not hopeless and still practically achievable through an iterative procedure as mentioned in Figure 1.1. The identifiability analyses can contribute to this procedure by improving the conditioning of the parameter estimation problem in different steps. First, *a priori* identifiability analysis can guide model refinement, (re)parametrization, and selection to produce models that are most succinct, i.e. not overly parameterized. Second, by restricting the parameter estimation only to search among parameters that are practically identifiable, the dimensionality of the parameter search space is reduced, allowing a faster convergence of the search algorithm to the global optima. Finally, the experiments for the next iteration can be designed to minimize the parameter confidence regions in the next cycle, for example using the so-called A-optimal design that minimizes the sum of the parameter variances [4, 128].

There is however a practical issue when using the *a priori* and two of the practical identifiability analyses (method 2 and 3) in the iterative loop. Since they rely on the sensitivity matrix as a linear approximation, this means that parameter values are needed to perform the analysis. Thus, the *a priori* identifiability can no longer be done prior to the experimentation or parameter estimation. In practice, the parameter values are initially guessed, or the identifiability analysis can be left out until the second iteration when parameter estimates are available from the first iterate. In this way, the identifiability analyses also iteratively improve along with the improvement of the model.

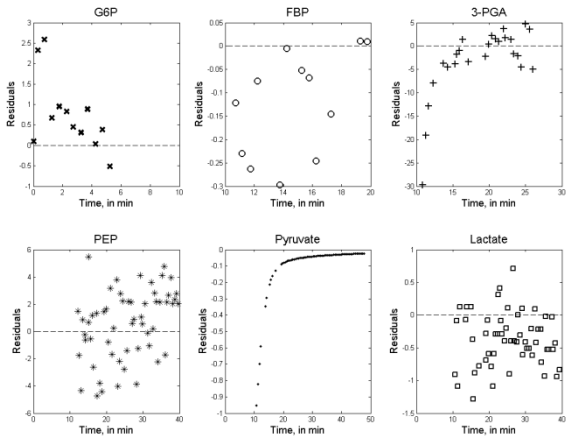

Figure 2.5. Residual Analysis of the *L. lactis* model

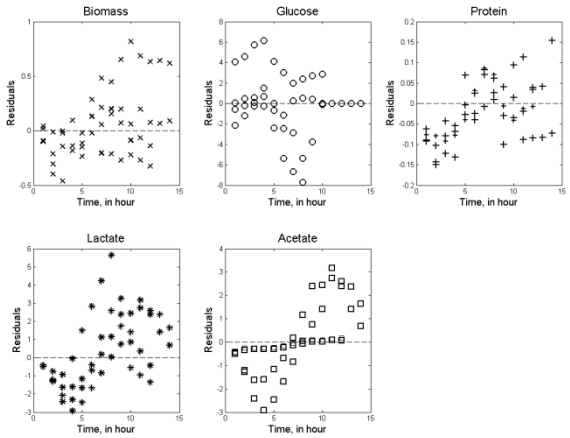

Figure 2.6. Residual Analysis of the *E. coli* model

Finally, the validity of the assumptions about the noise characteristics deserves to be discussed. From the plot of the residuals for the two models (see Figures 2.5 & 2.6), particularly in the case of *L. lactis* model, the assumption of constant variance is clearly violated, not to mention that the residuals do not appear to have zero mean nor Gaussian. Recall that in the derivation of the practical identifiability analyses, the model is implicitly assumed to be correct. Specifically in the BST, the reactions are assumed to

follow the power law formalism. In the *L. lactis* model, there is a clear sign of model inability to approximate the time-series measurements, and the model predictions are therefore biased, especially in early times. The variances used in the linearized identifiability analysis were consequently calculated from the residuals for $t > 10$ minutes, except in the case of G6P for which all of the residuals were accounted. Although this noise issue is less prominent in the second model (see Figure 2.6), such problem is likely to be more common than not.

Generally, Monte Carlo simulations can be used to characterize the contour of objective function hypersurface and define the confidence region exactly as in Eq.**(**2.18**)** [126, 137]. However, the high dimensionality of the parameters makes this approach impractical except for a small subset of the parameters. As mentioned previously, the iterative improvement of the model should also better the approximation of the parameter confidence regions. Thus, the practical identifiability analyses above should still be useful in the conditioning of the inverse modeling despite the validity of the noise assumption or the lack thereof.

2.4 CONCLUSIONS

Dynamic modeling of metabolic networks using BST models will lead to complicated optimization problems which in turn lead to multiple solutions and infeasible solutions. The inverse modeling problem in the BST is challenging even when using the state-of-the-art experimental data. This study shows that the root cause of the difficulty is the lack of parametric identifiability, which is affected by data quantity and quality (sampling rate, noise level), choice of experimental conditions (single or double

perturbations), and model parameterizations (S-systems or GMA). Applications to two example models in the BST showed that only about half of the parameters are *a priori* or structurally identifiable, and among these, only half are practically identifiable. This result suggests that most parameters cannot be uniquely and accurately identified from the data, regardless of the estimation algorithms used. This problem is perhaps more common than not in the inverse modeling of biological systems, and the parametric identifiability analyses can and should be integrated into the iterative procedure of biological modeling. Addressing these identifiability issues upfront can improve the subsequent reverse engineering of networks.

CHAPTER 3

IDENTIFIABILITY ANALYSIS OF DECOUPLED & LINLOG MODELS†

In this chapter, the methods developed in Chapter 2 are applied to analyze the alternative formalisms of BST, namely, decoupled system and linlog systems.

3.1 DECOUPLED MODELS

3.1.1 INTRODUCTION

The problem often faced in the parameter estimation of BST models is the difficulty in integration due to numerical stiffness, constituting almost 95% of time spent for the parameter searches [1]. Numerical integration of such dynamic models can be circumvented by fitting the differentials with slopes that are estimated from the time-series data at all measured points, essentially decoupling the ODE model [2]. The first half of this chapter focuses on the parameter identifiability analyses of such decoupled systems. The treatment of noise in the data will also be taken into account in detail. The analyses was applied to the decoupled versions of two previously published power-law models of metabolic networks: glycolytic pathway in *L. lactis* [60] and recombinant *E. coli* growth [132]. The results were then compared with that of the parameter estimation

† Excerpts of this work were published/presented in

1. Srinath, S.; Gunawan, R. In *Identifiability Analysis of Decoupled Power-Law Models*, 5th International Symposium on Design, Operation and Control of Chemical Processes (PSE Asia), Singapore, July 25-28, 2010; Singapore, 2010.
2. Srinath S, Gunawan R. Parameter Identifiability in Kinetic Modeling of Metabolic Pathways, In *Metabolic Engineering Conference VIII*, Jeju Island, South Korea, Jun 13 – 17, 2010

of the original ODE models presented in Chapter 2, revealing the differences between decoupled model and its corresponding BST model.

3.1.2 Mathematical Representation of Decoupled Models

Consider the ODE model given in Eq. (2.1). One parameter estimation method has been proposed that uses the derivatives $\dot{\mathbf{X}}$ at all measured time points t_k as slopes of the measured concentrations, in which parameter values are estimated by fitting each of $f_i(\mathbf{X};\boldsymbol{\theta})$ to the slope data. In essence, this method decouples the full ODE model identification into one ODE at a time. Figure 3.1 summarizes the procedure involved in decoupled estimation. In this case, the differential equations were not integrated in the estimation, alleviating the known ODE stiffness issue. Although decoupling the ODEs will result in a reduction of estimation time, the downside is the loss of complete mass balance among the metabolites across time, which often results in concentration predictions that offset the experimental data. In this case, by fitting only the concentration slopes, this method will only satisfy mass balance at each measurement time point.

3.1.3 Issues Related to Data

The datasets from biological measurements are usually noisy or incomplete. The incompleteness can be complemented by smoothing or standard interpolation. However, if a sufficiently large quantity of data or data from a particular time frame is missing, more complicated methods of prediction and bridging the gap are necessary. There are methods that can tackle the issue of missing data and although in reality there could be missing data, it is assumed not to be an issue here. So, the findings here are essentially for the best case scenario.

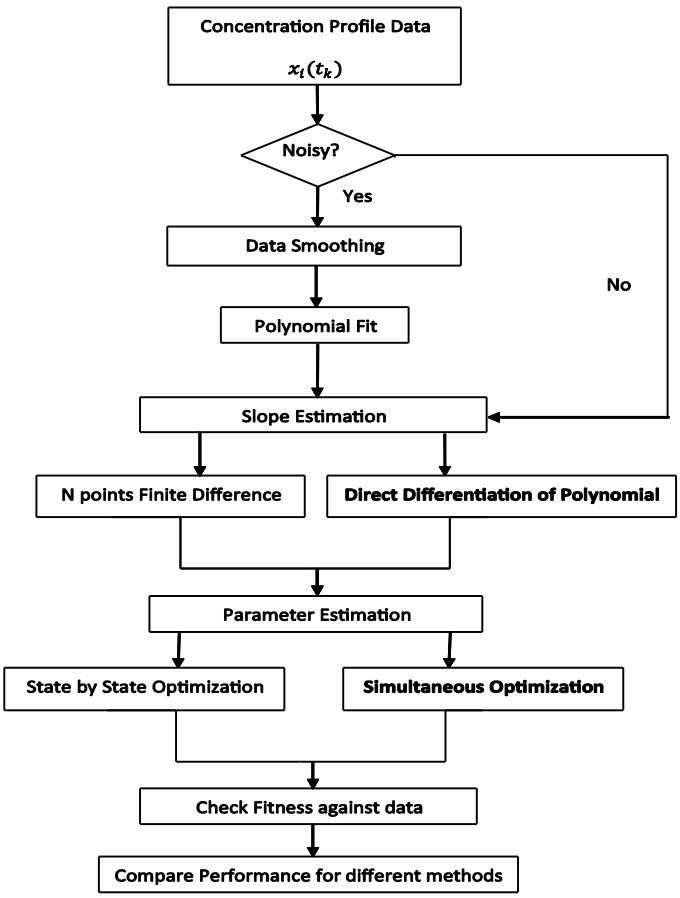

Figure 3.1. Flowchart of decoupled parameter estimation process.

When data are noisy, the slope estimation by traditional finite differences can give wildly fluctuating values, as random noises are amplified by such derivative calculations. Thus, data smoothing is necessary before slope calculations. Several smoothing techniques for the treatment of noisy data were proposed in literature, for example using B-splines smoothing [138] and artificial neural network (ANN) smoothing [2]. These smoothing techniques are proven to be effective when dealing with white noise. However, due to the complex smoothing algorithms involved, it is hard to trace the changes made to the data series and to estimate the variance associated with smoothing. Thus, it is important to develop a smoothing method that is unbiased, sufficiently simple

to allow tracking of changes and mathematical analysis, and yet can provide satisfactory smoothing performance that minimizes the effect of anomalous data points on parameter estimation. A good measure of determining whether a smoothing method is biased is to examine the residual (difference between the actual data and the estimated value) distribution.

3.1.4 Data Smoothing and Parameter Estimation

While the original publication of the decoupling method had used artificial neural network (ANN) model for data smoothing, other smoothing algorithms including moving average and polynomial fitting (polyfit) were investigated here. The main reason for choosing these algorithms over ANN is the necessity for tracking the propagation of the noise from the states to the parameters. Moving average (MA) is usually performed on time series data, for which the data are assumed to be "locally stationary" [139]. Depending on the weights assigned to different data points, the MA method has several variations: simple MA, weighted MA and exponentially weighted MA. In this work, a simple MA where equal weights are given to all data points is chosen. The simple moving average (SMA) can be calculated easily by taking the mean value of a fixed number of points. The SMA is able to smooth out local deviations, and reduce the effect of anomaly on the trend. Normally, the more data points used in MA, that is the larger the width of MA estimation, the smoother the data. However, if the width is too large that it becomes comparable to the width of the entire data series, the system dynamics cannot be fully captured. A preliminary study has been done (results not shown) and MA proved to be ineffective in alleviating the noise in the slope calculations and thus was not pursued further in this work (data not shown).

In general, polynomial fitting can be done in a piecewise manner and the polynomial fitting of concentrations ($\mathbf{X}_i$) versus time (t) can be written as:

$$\mathbf{X}^t = \mathbf{T}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} X^1 \\ X^2 \\ \\ X^{K-1} \\ X^K \end{bmatrix} = \begin{bmatrix} 1 & t_1 & \cdots & t_1^{p-1} & t_1^p \\ 1 & t_2 & \cdots & t_2^{p-1} & t_2^p \\ \vdots & \vdots & & & \\ 1 & t_{K-1} & \cdots & t_{K-1}^{p-1} & t_{K-1}^p \\ 1 & t_K & \cdots & t_K^{p-1} & t_K^p \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_{K-1} \\ \varepsilon_K \end{bmatrix} \tag{3.1}$$

where $\boldsymbol{\varepsilon}$ is the noise vector (zero mean, Gaussian, constant variance). In this work, the order of the polynomial was decided based upon the adjusted R^2 value, which takes account of the degrees of freedom. While an additional term in regression will always increase the traditional R^2 value, the adjusted R^2 value will only increase if the additional regressor improves data fitting [140]. In polynomial fitting, the adjusted R^2 value penalizes the use of higher order polynomials that do not improve data fit.

Using smoothened data, the time derivatives data $\mathbf{S}_i$ are directly obtained by differentiating the fitted polynomial equation and this is then followed by parameter estimation. Two approaches were used for optimization: (a) state by state (sequential) and (b) simultaneous (parallel). In the state by state approach, each and every state (X_i) is optimized sequentially, as given below.

$$\Phi_i = \min_{\boldsymbol{\theta}_i} \left(\mathbf{S}_i - f_i\left(\mathbf{X}, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_i\right)\right)^T \left(\mathbf{S}_i - f_i\left(\mathbf{X}, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_i\right)\right) \tag{3.2}$$

The parameter vector $\boldsymbol{\theta}_i$ is then passed into the next optimization involving state $\mathbf{X}_{i+1}$ with a similar objective function and by doing so the size of the parameter search space is much reduced. This process is then repeated until all measured states are accounted. Note

that the parameter values estimated from optimization of previous states are passed down and used directly in the subsequent optimization steps and thus, the resulting estimates obviously depend on the sequence of such optimizations. On the other hand, in the simultaneous estimation, the parameter values are estimated from a combined objective function, given by:

$$\Phi = \min_{\boldsymbol{\theta}} \sum_{j=1}^{n} \left(\mathbf{S}_j - f_j\left(\mathbf{X},\boldsymbol{\theta}\right)\right)^T \left(\mathbf{S}_j - f_j\left(\mathbf{X},\boldsymbol{\theta}\right)\right) \tag{3.3}$$

The comparison of these two approaches (results not shown) suggested that the simultaneous estimation, though more computationally expensive, gives better data fitting, as expected. Thus, in the subsequent sections, only the simultaneous estimation will be considered.

3.1.5 Identifiability Analysis

As mentioned in Chapter 2, model identifiability is defined as the ability to uniquely determine model structure and parameters from a given set of experimental data [98]. Using power-law formalism, model structure or connectivity among states can be inferred from the values of the kinetic order parameters and thus, model identifiability is equivalent to parameter identifiability [1]. To recap, there are two types of parameter identifiability; the first assumes noise-free data, referred to as *structural* or *a priori* identifiability, while the other accounts for random noise in data, referred to as *practical* identifiability.

In this chapter, data noise is again assumed to be independent and identically distributed (i.i.d.) random sample from a Gaussian distribution with zero mean and a

constant covariance matrix **V**. It is further assumed that the noise are uncorrelated giving a simple covariance matrix $\mathbf{V}=\sigma^2\mathbf{I}$. The propagation of noise is calculated in the following way. The variance associated with the coefficients **β** of the polynomial fitting can be expressed as

$$\mathbf{V}_\beta = \left(\sigma^2\left(\mathbf{T}^T\mathbf{T}\right)\right)^{-1} \tag{3.4}$$

Since **S** is calculated by direct differentiation of Eq. (3.1), the variance associated with the first derivative $\mathbf{V}_S$ can be expressed as

$$\mathbf{V}_S = \dot{\mathbf{T}}^T\mathbf{V}_b\dot{\mathbf{T}} \tag{3.5}$$

where $\dot{\mathbf{T}}$ is the first order derivative of **T** with respect to time. For the objective function in Eq. (3.3), the variance of the parameters (**θ**) can be approximated by:

$$\mathbf{V}_\theta = \left(\hat{\mathbf{F}}_\cdot^T\hat{\mathbf{F}}_\cdot\right)^{-1}\hat{\mathbf{F}}_\cdot^T\mathbf{V}_S\hat{\mathbf{F}}_\cdot\left(\hat{\mathbf{F}}_\cdot^T\hat{\mathbf{F}}_\cdot\right)^{-1} \tag{3.6}$$

where $\hat{\mathbf{F}}_\cdot$ is the sensitivity matrix, as defined in Eq. (2.15).

3.1.6 Results and Discussion

In this section, the parameter identifiability of the same two models as in Chapter 2 is presented, but with a key difference in the use of slope fitting here (instead of the concentration fitting in the previous chapter). Here, the parameters were obtained using the decoupled estimation and finally identifiability analyses were carried out for these estimated parameters. The quantity and quality of the data in these models arguably represent the best-case scenario of an inverse modeling problem in this area, where time-series concentrations of all metabolites within the subsystem are available and the model

dimensionality is small (number of states <10). All calculations were performed in MATLAB.

3.1.6.1 Case Study 1: Glycolytic pathway in *L. lactis*

In this work, noisy measurement data were generated by simulating the GMA model with the parameter values reported elsewhere [60]. Using these *in silico* data, a decoupled estimation is carried out, for which the data fitting is depicted in Figure 3.2. As mentioned above, the decoupled estimation often gives inaccurate estimates of the concentration profile since the data fitting is done using slopes. These estimated parameters were subsequently used to perform identifiability analyses. Using the actual experimental settings as done in the original publication [60], the total number of *a priori* identifiable parameters (AIP) was found to be 19 in total (7 rate constants and 12 kinetic orders). A summary of the *a priori* identifiability and practical identifiability is shown in Table 3.1. In using methods 2 and 3, the practical identifiability was performed only for AIP and based on a 95% confidence level. Method 1 however was applied for the complete parameters. According to Method 1, the number of identifiable parameters was only 13 (7 rate constants and 6 kinetic orders).

Table 3.1: Summary of identifiability results for both the models (Decoupled)

L. lactis* Model					
	Total Parameters	**AIP**	**Practical Identifiability**		
			Method 1	**Method 2**	**Method 3**
Rate constants	9	7	7	5	5
Kinetic order	16	12	6	7	7
***E. coli* Model#**					
Rate constants	10	9	7	8	9
Kinetic order	21	13	1	9	9

* At a sampling rate of **60** ***per hour***, a single bolus of Glucose

\# At a sampling rate of **1** ***per hour***, 40+50 g/L Glucose data

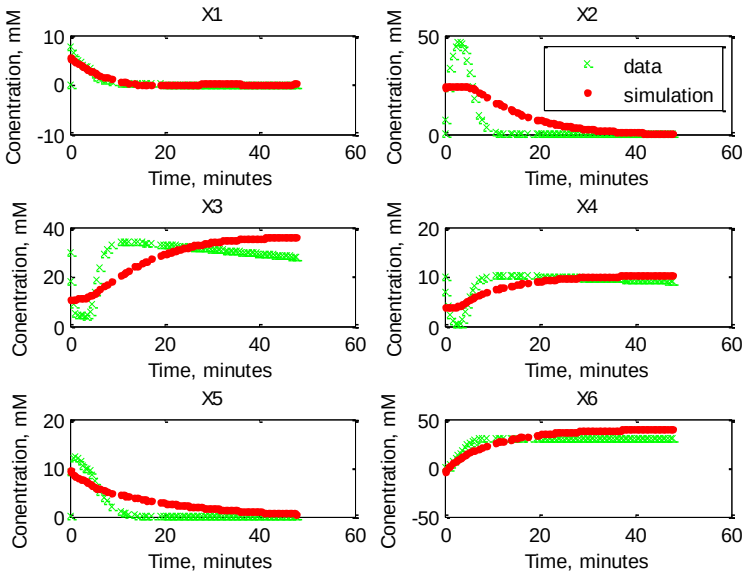

Figure 3.2. Comparison of *in silico* data and simulated profile for *L. lactis* model

3.1.6.2 Case Study 2: Modeling recombinant growth in *E. coli*

The second case study is the same S-system model of *E. coli* used in Chapter 2. The parameter estimates obtained by using decoupling approach were used for identifiability analysis. The data fitting results are shown in Figures 3.3 and 3.4 for the two different initial glucose concentrations. For these parameter estimates, 22 parameters were found to be *a priori* identifiable, comprising of 9 rate constants and 13 kinetic orders. The application of univariate practical identifiability analyses using 95% confidence level indicated that 18 out of 22 AIP (9 rate constants and 9 kinetic orders) were practically identifiable, and this number dropped to 17 (8 rate constants and 9 kinetic orders) based on the multivariate confidence region. The nonlinear multivariate method (Method 1) indicated that only 8 parameters (7 rate constants and 1 kinetic order) are practically identifiable. These results are also summarized in Table 3.1.

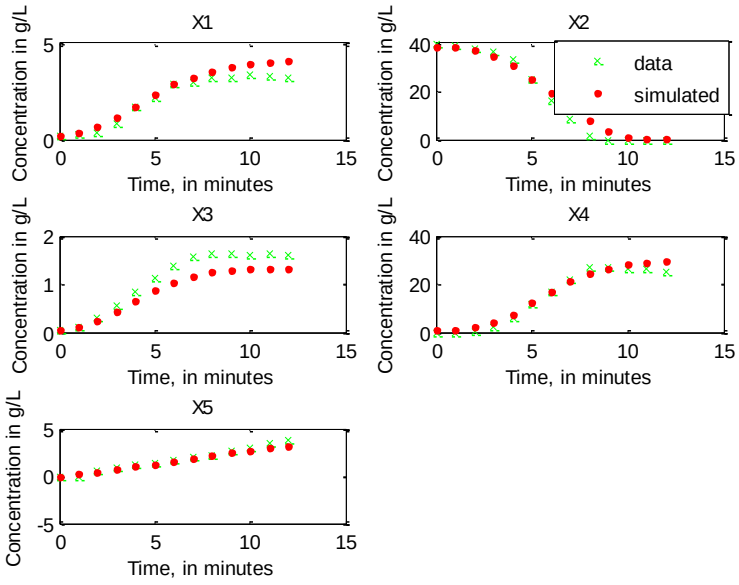

Figure 3.3. Comparison of *in silico* data and simulated profile for *E. coli* model for 40g/L Glucose concentration

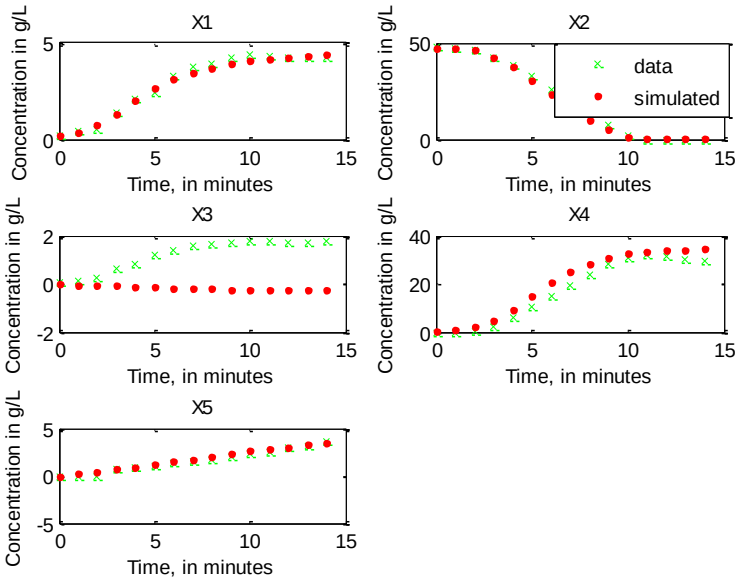

Figure 3.4. Comparison of *in silico* data and simulated profile for *E. coli* model for 50g/L Glucose concentration

3.1.6.3 Discussion

The results of the parametric identifiability study above confirm that the fundamental challenge faced in the inverse modeling of power-law models is rooted from the lack of complete parameter identifiability. This implies that the simultaneous estimation of all parameters from the experimental data is an ill-conditioned problem. Table 2.1 shows the summary of identifiability results for the parameter estimation using the ODE models. From Tables 2.1 and 3.1, it can be seen that decoupled estimation gave a higher number of identifiable parameters because decoupling results in lesser degree of correlation among states and parameters.

The decoupled estimation is formulated by converting the ODE model into algebraic equations and thus decoupling the states. While at first this may suggest that

these two models are the same, the comparison between identifiability results of the decoupled and direct ODE estimation indicated otherwise. The difference between the two models follows from some of the arguments presented in the earlier section. In the decoupled estimation, molar balance is only enforced at discrete time points, while in the ODE model, the balance is maintained at all time. As noted above, the decoupled estimation gives better identifiability property than ODE estimation, since the discrete time molar balance enforces a lesser constraint than continuous time. But, as shown in Figures 3.2, 3.3, and 3.4, the lack of continuous balance can lead to grossly inaccurate concentration profiles with offsets.

3.2 LINLOG MODELS

3.2.1 INTRODUCTION

Besides the power-law formalism, linear-logarithmic form has also been frequently used to develop kinetic models of metabolic networks [45]. As mentioned in Chapter 1 in power-law models, rates and variables are linearized in logarithmic spaces, and the models include the GMA and S-systems models. On the other hand, a linlog model is a hybrid linearization model, in which the rates are in linearly dependent on the logarithms of concentrations. There are two possible sub-types of linear-logarithmic models: (log)linear and linlog. A log-lin model is generally represented as $\ln(Z) = a + b(W)$ where the dependent variable (Z) is expressed on a logarithmic scale and the independent variable (W) is expressed in a linear metric. The logical inverse of the log-lin model is the linlog model. This model is represented as $Z = a + b\ln(W)$.

These models are closely associated with the MCA, an analytical approach for understanding the shared control within metabolic pathways close to a steady state.

3.2.2 LINLOG MODEL FORMALISM

A new approximative linear logarithmic (linlog) kinetic format was introduced by Hatzimanikatis [52] and recently expanded by Visser and Heijnen [53, 54]. This particular formalism was derived by combining a general kinetic model and theorems from the MCA. The model includes general expressions giving steady-state fluxes and metabolite levels as a function of enzyme levels, extracellular concentrations and the control and response coefficients. However, a limitation of linlog approximation is that the rate is undefined at zero metabolite concentration (x=0 or c=0).

Linlog models have some structural similarities with the BST [59]. The rates v_i in a linlog model are represented as

$$\frac{v_i}{J_i^0} = \frac{e_i}{e_i^0}\left(1 + \sum_{j=1}^{n+m} \varepsilon_{ij}^0 \ln\left(\frac{X_j}{X_j^0}\right)\right). \tag{3.7}$$

where J_i^0 is the *i*th reference steady-state flux rate, X_j^0 is the reference value of a dependent or independent variable, e_i^0 is the reference level of the *i*th enzyme activity and ε_{ij}^0 is the reference elasticity, which has the same interpretation as kinetic orders in BST. Eq. (1.7) and (3.7) are equivalent if the enzyme levels do not change during an experiment [141]. It is not always possible to determine the reference state of a dynamic system, so the reference variables are unknown or at least uncertain in parameter estimations. For the purposes of parameter estimation from time-series data, some of the

unknown parameters appear in fixed combinations and are merged into $\tilde{b}_i$ and $\tilde{a}_{ij}$ as shown in the following equations. The model of the metabolic network is formulated again as a system of differential equations in the following form:

$$v_i = -\tilde{b}_i + \sum_{j=1}^{n+m} \tilde{a}_{ij} \ln X_j . \tag{3.8}$$

where the coefficients $\tilde{a}_{ij}$ are equal to the product of the reference stead-state flux rate J_i^0 and the reference elasticities ε_{ij}^0 and the constant terms $\tilde{b}$ are given by

$$\tilde{b}_i = J_i^0 \left(\frac{e_i}{e_i^0} - \sum_{j=1}^{n+m} \varepsilon_{ij}^0 \ln X_j^0 \right). \tag{3.9}$$

The model has a nonlinear dependence on metabolite concentrations, but is entirely linear in its parameters. So, the parameter estimation of the linlog models is straightforward. The material balance equations are given by

$$\dot{X}_i = \sum_{j=1}^{r} S_{ij} v_j \tag{3.10}$$

where S_{ij} is the stoichiometric coefficient of the component *i* and reaction *j*. Substituting each linearized rate Eq. (3.8) into Eq. (3.10),

$$\dot{X}_i = -b_i + \sum_{j=1}^{n+m} a_{ij} \ln X_j \tag{3.11}$$

where b_i and a_{ij} are combination of $\tilde{b}_i$ and $\tilde{a}_{ij}$ with corresponding stoichiometric coefficients. As mentioned earlier, at steady-state, these equations are equivalent to an S-system model, with

$$a_{ij} = g_{ij} - h_{ij} \tag{3.12}$$

$$b_i = \ln\left(\frac{\beta_i}{\alpha_i}\right) \tag{3.13}$$

Similar equations can also be derived for a linlog model from GMA model [56].

3.2.3 Drawbacks of Linlog Models

The main limitation of the linlog model is that the model structure may not be able to capture the dynamics of the time courses adequately as pointed out by Voit and Chou [142]. The authors illustrated this drawback using a didactic branched pathway modeled originally by an S-system model. They have used the same model with different initial conditions and found that the dynamics of the estimated linlog model may be deviating from the data or lead to unreasonable results. The second and inherent drawback of linlog models is that rate is undefined at zero metabolite concentrations or sometimes leads to negative rates.

3.2.4 Results and Discussion

In this section, the parameter identifiability analysis of two linlog models: (1) *L. lactis* [56] and (2) *E. coli* metabolism [141], is presented. These two models are essentially the linlog version of the case studies discussed in Chapter 2. The linlog model equations of the *L. lactis* case study are given below [56].

$$
\begin{aligned}
\frac{dX_1}{dt} &= a_1 + a_2 \log Glc + \left(a_3 - a_5\right)\log X_1 + a_4 \log X_4 - a_6 \log ATP \\
\frac{dX_2}{dt} &= a_7 + a_5 \log X_1 + a_6 \log ATP - a_8 \log X_2 - a_9 \log P_i \\
\frac{dX_3}{dt} &= a_{10} + 2a_8 \log X_2 + 2a_9 \log P_i + a_{11} \log X_4 - a_{12} \log X_3 \\
\frac{dX_4}{dt} &= a_{13} + a_{12} \log X_3 - a_2 \log Glc - a_3 \log X_1 - a_{14} \log X_2 - \\
&\quad \left(a_4 + a_{11} + a_{15} + a_{17}\right)\log X_4 - a_{16} \log P_i \\
\frac{dX_5}{dt} &= a_{18} + a_2 \log Glc + a_3 \log X_1 + \left(a_{14} - a_{20}\right)\log X_2 + \\
&\quad \left(a_4 + a_{15}\right)\log X_4 - \left(a_{19} + a_{21}\right)\log X_5 + a_{16} \log P_i \\
\frac{dX_6}{dt} &= a_{22} + a_{19} \log X_5 + a_{20} \log X_2
\end{aligned}
\tag{3.14}
$$

The linlog parameters that appear as factors of terms with logarithms (non-constant terms in the linlog models) can be paired with corresponding kinetic orders of the GMA model [56] (for the GMA model refer Eq. (2.26)), which yields

$$
\begin{aligned}
&a_2 = g_{1,Glc}; a_3 = g_{11}; a_4 = g_{14}; a_5 = h_{11}; \\
&a_6 = h_{1,ATP}; a_8 = h_{22}; a_9 = h_{2,Pi}; a_{11} = g_{34}; \\
&a_{12} = h_{33}; a_{14} = h_{412}; a_{15} = h_{414}; a_{16} = h_{41,Pi}; \\
&a_{17} = h_{424}; a_{19} = h_{515}; a_{20} = h_{512}; a_{21} = h_{525};
\end{aligned}
\tag{3.15}
$$

For GMA equations containing only two terms, the S-systems method used in [141] was adopted. The corresponding parameters are

$$
a_1 = \frac{\log \beta_1}{\log \alpha_1}; \quad a_7 = \frac{\log \beta_2}{\log \beta_1}; \quad a_{22} = -\log \beta_{51} \tag{3.16}
$$

However, for GMA equations with more than two terms, such a direct comparison is not possible and hence Del-Rosario *et al* [56] suggested the following formulation:

$$
\begin{aligned}
a_{10} &= -\log\left(\frac{\beta_3}{2\alpha_3\beta_2}\right) \\
a_{13} &= -\log\left(\frac{\alpha_1\alpha_3\beta_{41}\beta_{42}}{\beta_3}\right) \\
a_{18} &= -\log\left(\frac{\beta_{51}\beta_{52}}{\alpha_1\beta_{41}}\right)
\end{aligned}
\tag{3.17}
$$

Using these expressions as starting guess values, all these parameters were estimated by Del Rosario *et al* [56]. The authors have used a series of parameter estimation methods to obtain reliable parameter estimates. The identifiability analyses here were carried out for the parameter estimates that were obtained from the simultaneous estimation method discussed in the aforementioned article.

For the second case study, the linlog model equations of the corresponding S-system (given in Eq. (2.27)), are given by:

$$
\begin{aligned}
\frac{dX_1}{dt} &= -b_1 + a_{11}\log X_1 + a_{12}\log X_2 \\
\frac{dX_2}{dt} &= -b_2 + a_{21}\log X_1 + a_{22}\log X_2 \\
\frac{dX_3}{dt} &= -b_3 + a_{31}\log X_1 + a_{32}\log X_2 + a_{33}\log X_3 + a_{34}\log X_4 \\
\frac{dX_4}{dt} &= -b_4 + a_{41}\log X_1 + a_{42}\log X_2 + a_{44}\log X_4 \\
\frac{dX_5}{dt} &= -b_5 + a_{51}\log X_1 + a_{52}\log X_2 . + + a_{55}\log X_5
\end{aligned}
\tag{3.18}
$$

The parameters estimated by Wang *et al* [141] was used for parameter identifiability of this system.

Table 3.2: Summary linlog parameter identifiability results

L. lactis* Model					
	Total Parameters	**AIP**	**Practical Identifiability**		
			Method 1	**Method 2**	**Method 3**
Rate constants	6	3	0	0	0
Kinetic order	16	12	3	0	0
***E. coli* Model#**					
Rate constants	5	3	0	1	1
Kinetic order	14	7	1	2	3

* At a sampling rate of **60** ***per hour***, a single bolus of Glucose

\# At a sampling rate of **1** ***per hour***, 40+50 g/L Glucose data

The summary of identifiability results is presented in Table 3.2. It is clear from this table that the identifiability of the linlog models (Eq. (3.14) & (3.18)) is poor compared to their original power-law model (Table 2.1), even though the linlog rates are linear in their parameters. The poor performance is due to the interplay of the following two reasons. Firstly, as mentioned, the structure of linlog model is not suited for data sets where variables assume values close to zero. Wang *et al* [141] clearly showed that emergence of negative rates in the linlog models is not limited to irrelevantly small concentrations. Secondly, another problem could arise from not achieving the global optima in the parameter estimation. In order to distinguish between the last two possibilities, Del Rosario *et al* [56] performed parameter estimation using only transient

data and used linlog system to fit a small model of gene regulatory network. As observed in that work, the linlog model was able to fit the transient data but encountered problems with negative concentration predictions or stiffness when running the model for longer time periods. Although the parameter estimation of linlog models is a straightforward task, the results here suggest that the estimated parameters are reliable only if they are estimated from data that do not involve near zero concentration values. The occurrence of negative rates does not imply that linlog models are necessarily bad; it simply means that these models are suitable in some, but not all situations. The decision of choosing a model structure should depend on the operating ranges of the concentration that are expected in the modeled experiment. If these ranges stretch from normal to large, then linlog model might be suitable. On the other hand if the concentrations are quite small, the linlog model might be in accurate [141].

3.3 CONCLUSIONS

The major motivation of using decoupled parameter estimation is the fact that simulation of ODE models consume significant amount of time. The parameter identifiability study of two examples from the BST gave a different set of identifiable parameters for the ODE and decoupled estimations, thereby suggesting that the two underlying models are different. The decoupling of the ODEs resulted in a significant reduction of computation time at the cost of losing out on partial molar balance among the metabolites which led to poor prediction of the metabolite concentrations. The identifiability analysis also suggested that most parameters are still not practically identifiable from data, regardless of the estimation used.

Linlog model is a relatively new mathematical framework that combines a general kinetic model and theorems from MCA. This framework includes general expressions giving steady-state fluxes and metabolite levels as a function of enzyme levels, extracellular concentration and the control and the response coefficients. The number of parameters in linlog model is minimal and all rate equations have the same mathematical structure as in the BST models. Parameter estimation of the BST models proved to be a bottleneck and hence linlog models are a good alternative as they simplify the parameter estimation task into linear regression. The parameter identifiability of the linlog models was poor for the same two examples since the dynamics of input substrate (glucose) approached zero, a situation that is quite common in bolus and single-batch experiments [141]. The results here cannot be generalized to all linlog models because if the experiments are designed at far from zero concentrations the parameter identifiability could be certainly improved. As mentioned previously, linlog models are developed to describe dynamics near the steady state, but the single-batch experiment (Case Study 2) is clearly not at steady state. This could also be a reason for poor identifiability results for Case study 2.

CHAPTER 4†

DESIGN OF EXPERIMENTS

4.1 INTRODUCTION

Design of experiments (DOE) is often necessary to achieve predictive knowledge of a complex, multivariate process with fewest trials possible. This concept is quite important because the collection of experimental data is costly and requires careful planning such that maximum information can be achieved at minimum resource utilization. One of the most commonly used designs is the factorial design, full or fractional, which allows the simultaneous examination of the effects multiple independent variables and their degree of interactions. When the limitations of time and resources prevent the experimental exploration of all feasible behavior in a certain process, the use of mathematical model could overcome this drawback. This is called as Model-Based Design of Experiments (MBDOE).

Experimental data is collected for several reasons: (1) to get a better understanding of some phenomena of the system under study, (2) to estimate the parameters of a model or (3) to discriminate between several possible model structures. The focus of this chapter will be on optimal experiment design for parameter estimation. The next two sections discuss about MBDOE, providing a brief overview of the dynamic optimization framework for the DOE problem and highlighting the drawback in the most

† Excerpts of this work was submitted to
Srinath S and Gunawan R. *Multiobjective Optimization of Experiments: Curvature and Fisher Information Matrix*, AiChe, *In Review*

commonly used MBDOE based on the linearization of the model. The subsequent section presents curvature-based MBDOE wherein nonlinearity of the model is taken into account and also discusses about the proposed multiobjective optimization (MOO) of a design criterion. This method is then applied to a couple of examples and its performance is compared with those of other design criteria.

4.2 MODEL-BASED DESIGN OF EXPERIMENTS

MBDOE offers a means for combining modeling and experimental efforts such that the knowledge generated from prior experimental data and modeling effort, as contained in the model equations and parameters, is used to guide subsequent experiments. MBDOE is one step of and closes the loop in the iterative procedure for model identification (see Figure 1.1). In particular MBDOE aims at aiding the experimenter in devising experimental strategies that would give maximum information for estimating the unknown parameters to high precision. MBDOE techniques are extensively used in process and systems engineering and an excellent review of MBDOE can be found in [143]. Smith [144] was one of the first researchers to state a design criterion and obtain optimal experiment designs for regression problems which was later termed as G-optimality (which minimizes the maximum of the standardized variance over the design space, as defined by Kiefer and Wolfowitz [145]). Although the work on theory of optimal design for linear model was initiated in 1918, nonlinear models were not considered until 1959 by Box and Lucas [146].

In general, noisy measurement data $\mathbf{y} \in \mathbb{R}^n$ can be described in the following manner:

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \tag{4.1}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\varepsilon}$ denote the measurement mean and random noise, respectively. In a typical nonlinear regression problem, the total number of data points *n* is usually much larger than the number of parameters *p* and consequently, $\boldsymbol{\mu}$ spans a *p*-dimensional space $\Omega \subset \mathbb{R}^n$, where

$$\Omega = \left\{ \boldsymbol{\mu} : \boldsymbol{\mu} = \mathbf{F}\left(\mathbf{x}, \mathbf{u}, \boldsymbol{\theta}\right), \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p \right\}. \tag{4.2}$$

Here, **x** denotes the state vector, **θ** is the parameter vector, **u** is the input, and $\mathbf{F}\left(\mathbf{x}, \mathbf{u}, \boldsymbol{\theta}\right)$ is a general vector-valued nonlinear equation. The subspace Ω is also called the *expectation surface* or the *solution locus.*

Consider the ODE model described in Eq. (2.1). Now, taking the Taylor series expansion of **f**(**θ**;**x**) around the nominal parameters $\hat{\boldsymbol{\theta}}$,

$$\begin{aligned} \mathbf{y} &= \mathbf{f}\left(\hat{\boldsymbol{\theta}}; \mathbf{x}\right) + \left.\frac{\partial \mathbf{y}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}^*; \mathbf{x}} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) + \ldots \\ \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} &\approx \left.\frac{\partial \mathbf{y}}{\partial \boldsymbol{\theta}}\right|^{-1}_{\boldsymbol{\theta}^*; \mathbf{x}} \left(\mathbf{y} - \mathbf{f}\left(\hat{\boldsymbol{\theta}}; \mathbf{x}\right)\right) = \hat{\mathbf{F}}_{\cdot}^{-1}\left(\hat{\boldsymbol{\theta}}; \mathbf{x}\right)\left(\mathbf{y} - \mathbf{f}\left(\hat{\boldsymbol{\theta}}; \mathbf{x}\right)\right). \end{aligned} \tag{4.3}$$

where $\hat{\mathbf{F}}_{\cdot} = \mathbf{F}_{\cdot}\left(\hat{\boldsymbol{\theta}}; \mathbf{x}\right)$ is the first order sensitivity or the Jacobian matrix, i.e. $\left.\frac{\partial \mathbf{F}\left(\mathbf{x}, \mathbf{u}, \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}}$ as defined in Eq. (2.15).

The estimation of model parameters **θ** from a given set of measurement data **y** is typically formulated as a minimization of the weighted sum of squares of the difference between model prediction $\mathbf{F}\left(\mathbf{x}, \mathbf{u}, \boldsymbol{\theta}\right)$ and the data **y**. If and when the noise is normally distributed with a constant variance **V**, one special case of the above minimization

problem is known as the maximum likelihood estimation, in which the model parameters are estimated by minimizing the following objective function:

$$\Phi(\boldsymbol{\theta}) = \left(\mathbf{y} - \mathbf{F}(\mathbf{x},\mathbf{u},\boldsymbol{\theta})\right)^T \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{F}(\mathbf{x},\mathbf{u},\boldsymbol{\theta})\right)$$

Furthermore, if the model is linear, i.e. $\mathbf{F}(\mathbf{x},\mathbf{u},\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\theta}$, the least square parameter estimates are given by $\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}$, which is also the minimum variance unbiased estimator with the minimum variance of $\mathbf{V}_\theta = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$. In a general nonlinear model, the minimum of $\Phi(\boldsymbol{\theta})$ however does not necessarily correspond to the minimum variance estimator of $\boldsymbol{\theta}$, but using the Cramer-Rao inequality [147], the minimum variance of the estimate $\hat{\boldsymbol{\theta}}$ can be approximated using the inverse of the Fisher information matrix (FIM), given by

$$\mathbf{V}_\theta \geq \mathrm{FIM}^{-1} = \left(\hat{\mathbf{F}}_\bullet^T \mathbf{V}^{-1} \hat{\mathbf{F}}_\bullet\right)^{-1} \tag{4.4}$$

where $\mathbf{V}$ is a non-negative definite symmetric matrix which is related to the experimental error. It is worth noting that Eq. (4.4) is valid only asymptotically for nonlinear models. The necessary conditions are that the measurement noise is uncorrelated and has zero mean. It is also required that the residuals are uncorrelated and white. It is worth noting that for nonlinear models Eq. (4.4) holds well only asymptotically [148].

The pioneering works done in the field of MBDOE were contributed by Wald [149], Chernhoff [150], Ehrenfeld [151], Box and Lucas [146] and Kiefer and Wolfowitz [145]. All of these works considered mainly steady-state models (both linear and nonlinear). Extension of DOE to dynamic system was a slow process [152], but the

potential benefits of DOE for dynamic systems was already recognized in 1977 by Goodwin and Payne [153]. Since then there have been umpteen numbers of works that have successfully applied the MBDOE technique to various systems in various fields [128, 131, 154-157].

4.3 DYNAMIC OPTIMIZATION FRAMEWORK

In MBDOE, the design of experiments is typically casted as an optimization problem. In order to do so, the experimental conditions will need to be first parameterized. It has to be decided when and how to perturb the system, when and how the measurements will be performed on the system. Depending on the experimental setting, these design variables might have simple bounds and/or optional nonlinear equality/inequality constraints on the initial conditions, response variables and state variables [158]. All these quantities together form design constraints. In the following, the problem of designing dynamic experiments is formulated as a control vector parameterization (CVP), which enables the calculation of fixed number of optimal sampling points, experimental duration and initial conditions of the experiment. The time-varying inputs are approximated as a piecewise constant.

$$\begin{aligned} u_i(t) &= z_{ij} \qquad \forall t \in \tau_{sw,j} \\ i &= 1,2...,n_c \quad j = 1,2,...,n_{sw} \end{aligned} \tag{4.5}$$

where n_c is the number of time varying controls and n_{sw} is the number of switching intervals, $\boldsymbol{\tau}_{sw}$ defines the intervals in which the time varying controls are constant at $z_{i,j}$. So, by using CVP, the time varying input trajectories are represented with a finite number of optimization variables (see Figure 4.1). Apart from the inputs, the initial conditions of

the experiment $\mathbf{x}_0$ can also be considered as design variables. In this chapter, the design variables include the initial condition of the states ($\mathbf{x}_0$), the time points of measurements ($\mathbf{t}_{sp}$) and the piecewise-constant dynamic input (see Figure 4.2), which was parameterized by the magnitudes ($\mathbf{u}$) and time steps ($\mathbf{t}_{sw}$). The total number of measurement time points and the input time steps will be prescribed prior to the design.

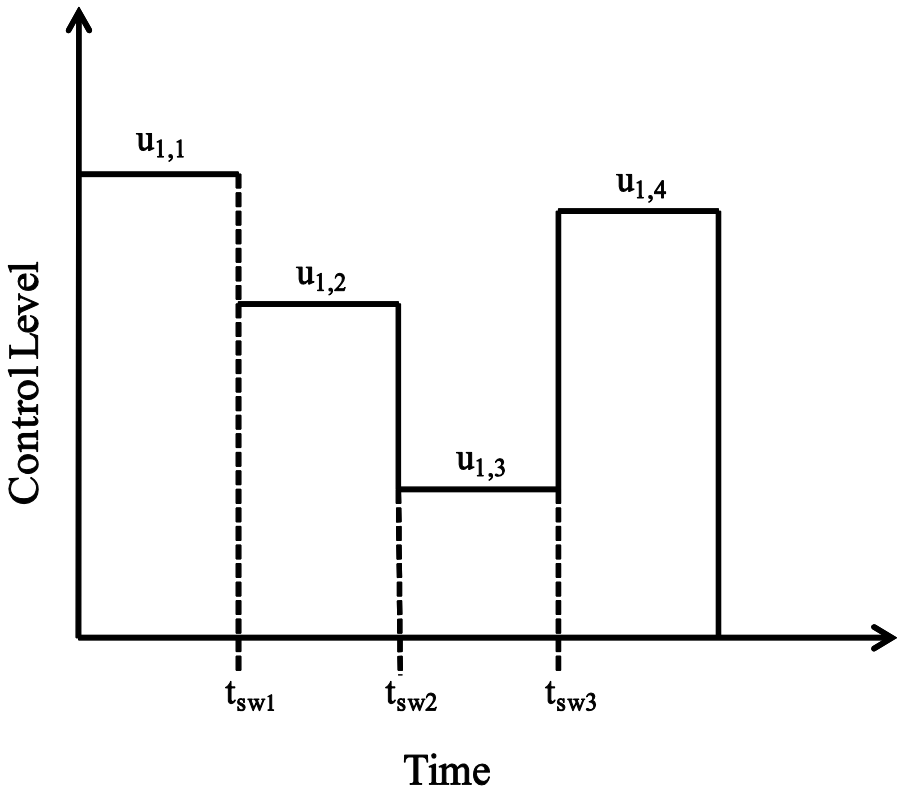

Figure 4.1. Illustration of piecewise constant input profile

Before initiating the search procedure for optimal experiment design, an objective has to be defined. As the FIM is inversely related to the parameter variance (Eq. (4.4)), the MBDOE is typically casted as a maximization of some metric of the FIM or minimization of a measure of its inverse, e.g. by using eigenvalues and eigenvectors of the FIM. Specifically, the eigenvalues of the FIM are inversely proportional to the size of

the axes of the parametric confidence (hyper)ellipsoids, i.e. the larger eigenvalues correspond to the smaller confidence ellipsoid axes and vice-versa (see Figure 4.2). Different scalar properties based on the eigenvalues of the FIM are used as metrics for optimal experiment design. Some of the popular scalar measures are described in Table 4.1.

Table 4.1: FIM-based design of experiment criteria

FIM-based MBDOE	**Criterion***
A-optimal	$\max \sum_i \lambda_i$
D-optimal	$\max \prod_i \lambda_i$
E-optimal	$\max \lambda_{\min}$
Modified E-optimal	$\max \left(\lambda_{\min} / \lambda_{\max} \right)$

* λ_i's are eigenvalues of FIM

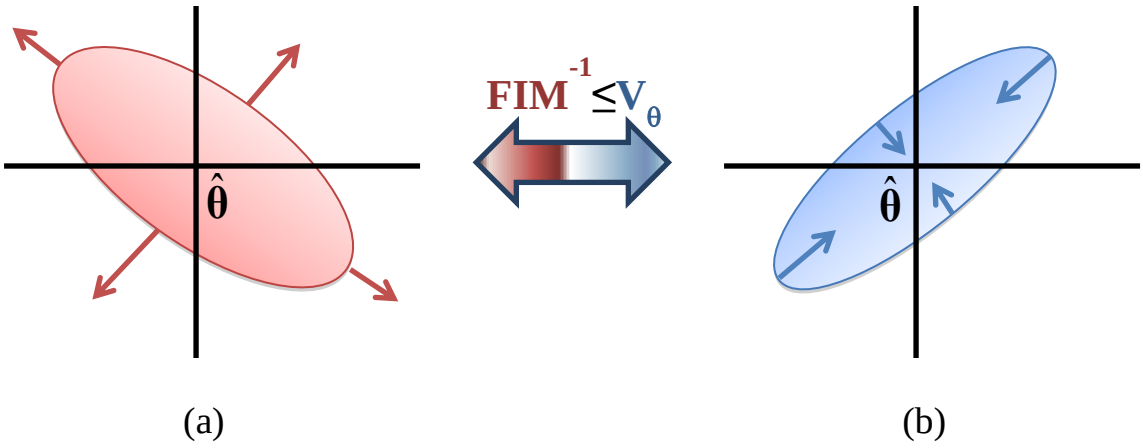

Figure 4.2. Two parametric confidence ellipse. The axes represent the two parameters and origin represents the parameter estimate (a) Ellipse of information. The axes of this ellipse are characterized by the eigenvalues of the FIM. D-optimality maximizes the volume of this ellipse (as indicated by the arrows) (b) Ellipse of uncertainty. The axes of this ellipse are defined by the inverse of the eigenvalues of FIM. A-optimality minimizes this region of uncertainty (as indicated by the arrows).

The main weakness of FIM-based MBDOEs is the underlying linear model assumption, which is done here through the use of a linearized output of the model, i.e. the Jacobian. Essentially, this linearization replaces the expectation surface Ω by its tangent plane at $\hat{\boldsymbol{\theta}}$. By doing so, two approximations have been made in the design of experiment; (1) that the model outputs vary proportionally with the parameter values (planar assumption) and (2) that this proportionality is constant (uniform coordinate assumption) [87]. If the model is highly nonlinear, such MBDOE based on its linearization could be severely suboptimal [159, 160].

4.4 CURVATURE BASED DESIGN OF EXPERIMENTS

Cochran [161] first noted the asymptotic nature of FIM-based design and invited studies of its small sample performance. In its response, Box [162] derived an approximation for the bias of the least square estimators and suggested designing experiments to minimize this bias. Clarke [163] derived an improved formula for the variance-covariance matrix of the parameter estimator by considering a term beyond the usual linear approximation and recommended designing experiments to minimize the mean squared error of the estimator. Bates and Watts [164] proposed selecting designs to minimize the curvatures of the model as measured by Hessian matrix to simplify the inference procedures. Hamilton and Watts [165] developed a quadratic design criterion, also called as Q-optimality [159], based on second order approximation of the volume of parameter inference region. As the validity of the linear approximation quite strongly depends on the nonlinearity of the model function, it is sensible to utilize the curvature measures developed by Bates and Watts [164] in developing a curvature-based optimum design criterion.

Curvature-based designs of experiment have been introduced to mitigate the linearization issue related to the FIM, by using a second order approximation of the (nonlinear) model output. In particular, the curvature of the expectation surface is captured using the second order sensitivities of $\mathbf{F}(\mathbf{x},\mathbf{u},\boldsymbol{\theta})$ from the following Taylor series expansion:

$$\mathbf{F}(\mathbf{x},\mathbf{u},\boldsymbol{\theta}) = \mathbf{F}\left(\mathbf{x},\mathbf{u},\hat{\boldsymbol{\theta}}\right) + \hat{\mathbf{F}}_{\cdot}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right) + \frac{1}{2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^{T}\hat{\mathbf{F}}_{\cdot\cdot}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right) + \mathrm{O}\left(\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^{3}\right) \tag{4.6}$$

As it can be seen, the quadratic approximation involves the calculation of first ($\hat{\mathbf{F}}_{\cdot}$) and second order sensitivities ($\hat{\mathbf{F}}_{\cdot\cdot}$). To determine the validity of the linear assumption, the second order derivatives of the expectation surface are used to derive curvature measures. There exist several design strategies based on such a quadratic model approximation. For example, Hamilton and Watts [165] had previously derived a design criterion, called the Q-optimality [159], that minimizes the second order approximation of the volume of parameter confidence region. This method was a natural extension of the FIM-based D-optimality design, and had been further refined [166] and applied to steady state model identification [158, 159, 167]. In addition, Benabbas *et al* [158] had used the Hessian matrix directly as an indicator of model nonlinearity to optimize experiments for parameter estimation in dynamical systems. In their two curvature-based designs of experiments, the root mean square (RMS) of the elements in this matrix was either minimized or guaranteed to be lower than a certain acceptable level, of which the latter was the better performing.

4.4.1 Multi- Objective Design of Experiment

The existing curvature-based designs either minimized only the curvature or constrained the curvature to threshold value and maximized information. The problem with the former approach is that the information (FIM) is not maximized and the latter approach has a drawback of choosing a threshold value for the curvature which is arbitrary and case-dependent. To counter these problems, in this work, a MOO approach is taken in which information is maximized and simultaneously the curvatures are minimized. The basic premise of this method is to select experimental conditions where

model outputs can be sufficiently described by its linearization (i.e. Jacobian) and at the same time the informativeness of data is maximized. Hence, two of the three objective functions in the formulation below describe the minimization of the relative curvature measures[164], while the last objective function is equivalent to the FIM based D-optimal design.

In deriving the relative curvature measures, consider first an arbitrary straight line in the parameter space passing through $\hat{\boldsymbol{\theta}}$:

$$\boldsymbol{\theta}(b) = \hat{\boldsymbol{\theta}} + b\mathbf{h} = \hat{\boldsymbol{\theta}} + \boldsymbol{\delta} \tag{4.7}$$

where $\mathbf{h} = (h_1, ..., h_p)$ is any non-zero vector. As the scalar parameter b is varied, a curve traces through the expectation surface, which is also referred to as a lifted line, according to:

$$\boldsymbol{\mu}_{\mathbf{h}}(\boldsymbol{\theta}) = \boldsymbol{\mu}\left(\hat{\boldsymbol{\theta}} + b\mathbf{h}\right). \tag{4.8}$$

Since $\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{f}(\boldsymbol{\theta})$, the tangent to this curve (CD in the Figure 4.3) at b=0 is

$$\begin{aligned} \dot{\boldsymbol{\mu}}_{\mathbf{h}} &= \left[\frac{d\boldsymbol{\mu}_{\mathbf{h}}(\boldsymbol{\theta})}{db}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, b=0} \\ &= \left[\sum_{r=1}^{p} \frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \theta_r} \frac{\partial \theta_r(b)}{\partial b}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, b=0} \\ &= \hat{\mathbf{F}}\mathbf{h}. \end{aligned} \tag{4.9}$$

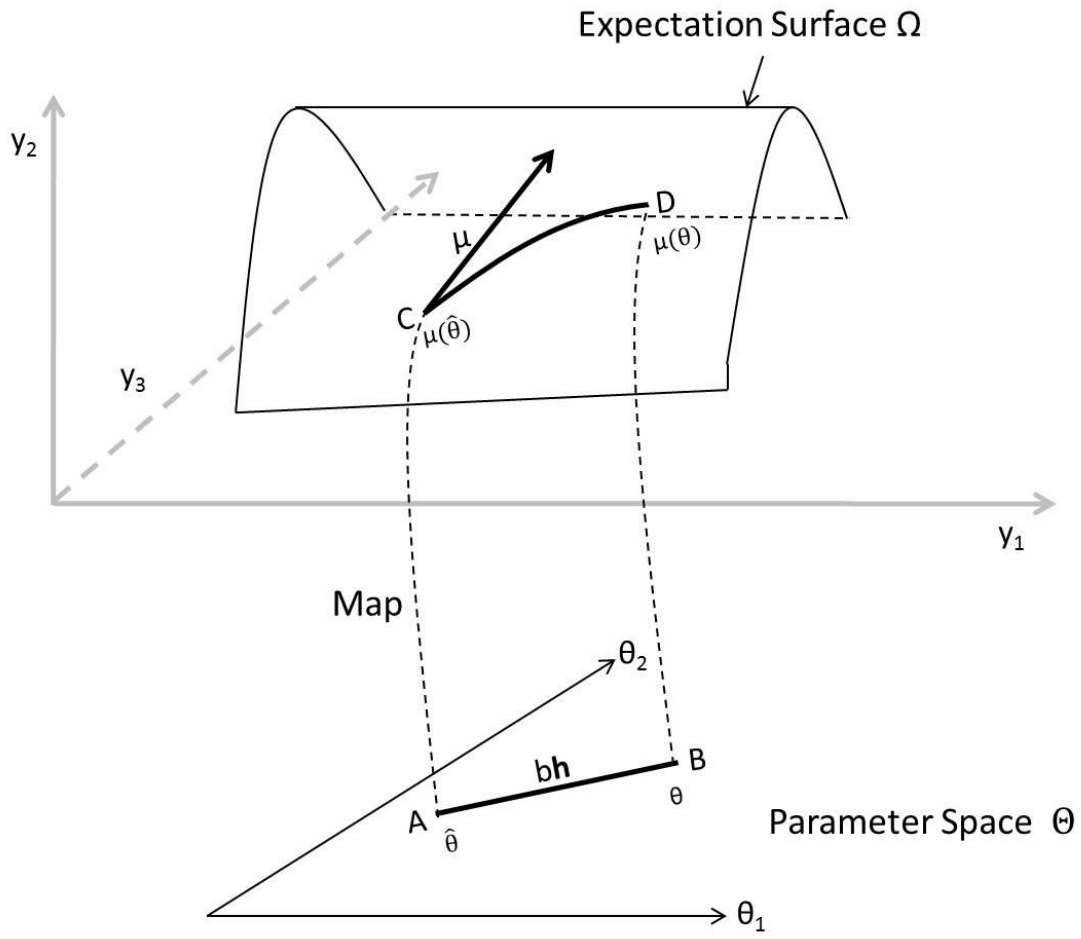

Figure 4.3. Expectation surface and Parameter space (*Adapted from Seber and Wild* [87])

The set of all such tangent lines, i.e. the column space of $\hat{\mathbf{F}}_{\cdot}$, gives the tangent plane at $\boldsymbol{\mu}\left(\hat{\boldsymbol{\theta}}\right)$. The curvature measures rely on the quadratic approximation of $\boldsymbol{\mu}$ and in this case, one can write the acceleration of $\boldsymbol{\mu}_{\mathbf{h}}\left(b\right)$ at $b = 0$ as follows:

$$\ddot{\boldsymbol{\mu}}_{\mathbf{h}} = \mathbf{h}^{T}\hat{\mathbf{F}}_{\cdot\cdot}\mathbf{h} = \sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\partial^{2}\mathbf{F}\left(\mathbf{x},\mathbf{u},\boldsymbol{\theta}\right)}{\partial\theta_{i}\partial\theta_{j}}h_{i}h_{j}. \tag{4.10}$$

The acceleration vector $\ddot{\boldsymbol{\mu}}_{\mathbf{h}}$ can be subsequently decomposed into two components; $\ddot{\boldsymbol{\mu}}_{\mathbf{h}}^{n}$, which is normal to the tangent plane, and $\ddot{\boldsymbol{\mu}}_{\mathbf{h}}^{t}$, which is tangential to the tangent plane at $\boldsymbol{\mu}\left(\hat{\boldsymbol{\theta}}\right)$, i.e.

$$\ddot{\boldsymbol{\mu}}_{\mathbf{h}} = \ddot{\boldsymbol{\mu}}_{\mathbf{h}}^{t} + \ddot{\boldsymbol{\mu}}_{\mathbf{h}}^{n}. \tag{4.11}$$

Physically, the normal component determines the change in direction of the vector $\dot{\boldsymbol{\mu}}_{\mathbf{h}}$ *normal* to tangent plane and the tangential component determines the speed of the moving point and hence determines whether the point moves uniformly across the solution locus. The tangential acceleration is also called the parameter-effects curvature [164] as it provides a measure of nonlinearity along the parameter vector **h**. While the degree of the parameter-effects curvature can be adjusted through (re)parameterization of model equations, the normal acceleration does not change with model parameterization, and hence is named the intrinsic curvature. Finally, the relative curvature measures in the direction of **h** are given by [87, 164]

$$K_{\mathbf{h}}^{\mathbf{t}} = \frac{\left\| \ddot{\boldsymbol{\mu}}_{\mathbf{h}}^{t} \right\|}{\left\| \dot{\boldsymbol{\mu}}_{\mathbf{h}}^{t} \right\|^{2}} \tag{4.12}$$

$$K_{\mathbf{h}}^{n} = \frac{\left\| \ddot{\boldsymbol{\mu}}_{\mathbf{h}}^{n} \right\|}{\left\| \dot{\boldsymbol{\mu}}_{\mathbf{h}}^{n} \right\|^{2}} \tag{4.13}$$

The decomposition of the Hessian into the tangential and the normal component is described below.

Since it is desired to determine the lengths of the components $\ddot{\boldsymbol{\mu}}_{\mathbf{h}}^{n}$ and $\ddot{\boldsymbol{\mu}}_{\mathbf{h}}^{t}$ the coordinates of sample space are rotated so that the first *p* coordinate vectors are parallel

to the tangent plane and the last *n-p* are orthogonal to it. This can be accomplished by pre-multiplying all the vectors in sample space by an orthogonal matrix $\mathbf{Q}'$, where $\mathbf{Q}$ is part of the **QR** decomposition of $\hat{\mathbf{F}}_{\bullet}$. That is

$$\hat{\mathbf{F}}_{\bullet} = \mathbf{QR} = \mathbf{Q}\begin{bmatrix}\tilde{\mathbf{R}}\\ 0\end{bmatrix}. \tag{4.14}$$

where $\tilde{\mathbf{R}}$ is upper triangular. By rotating the parameter axes $\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)$ into $\boldsymbol{\varphi}=\tilde{\mathbf{R}}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)$, a new Jacobian matrix $\mathbf{U}_{\bullet} = \left.\dfrac{d\mathbf{F}\left(\mathbf{x},\mathbf{u},\boldsymbol{\varphi}\right)}{d\boldsymbol{\varphi}}\right|_{\boldsymbol{\varphi}=0}$ can be computed as $\mathbf{U}_{\bullet} = \hat{\mathbf{F}}_{\bullet}\tilde{\mathbf{R}}^{-1}$, which comprises the first *p* column vectors of $\mathbf{Q}$ (i.e. $\mathbf{Q}=\left[\mathbf{U}_{\bullet} \quad \mathbf{N}\right]$). The remaining column vectors of $\mathbf{Q}$ (i.e. $\mathbf{N}$) are orthonormal to the tangent surface at $\boldsymbol{\varphi}=0$. In the same manner, the Hessian matrix in the rotated axes can be written as $\mathbf{U}_{\bullet\bullet} = \mathbf{L}^{T}\hat{\mathbf{F}}_{\bullet\bullet}\mathbf{L}$, where $\mathbf{L}=\tilde{\mathbf{R}}^{-1}$ and $U_{\bullet\bullet ijk} = \left.\dfrac{\partial^2 F_i\left(\mathbf{x},\mathbf{u},\boldsymbol{\varphi}\right)}{\partial\varphi_j\partial\varphi_k}\right|_{\boldsymbol{\varphi}=0}$ and the decomposition of the Hessian into the tangential and normal component is given by the following equation:

$$\mathbf{A}_{\bullet\bullet} = \mathbf{Q}^{T}\mathbf{U}_{\bullet\bullet} = \left[\mathbf{U}_{\bullet} \quad \mathbf{N}\right]^{T}\mathbf{U}_{\bullet\bullet} = \left[\mathbf{A}^{t}_{\bullet\bullet} \quad \mathbf{A}^{n}_{\bullet\bullet}\right] \tag{4.15}$$

In this case, $\mathbf{A}^{t}_{\bullet\bullet}$ and $\mathbf{A}^{n}_{\bullet\bullet}$ are the parametric and intrinsic curvature component of the Hessian, respectively. The normal component, intrinsic curvature array, measures the degree of nonlinearity inherent to the model itself. The tangential component, parameter-effects curvature array, measures the degree of nonlinearity depending on the parameterization in the model. The intrinsic nonlinearity does not depend on the model

parametrization but only on the experimental design and the expression for the expectation surface.

To illustrate the concepts of intrinsic nonlinearity and the parameter effects nonlinearity, an example taken from Bates and Wild (1988) [168] is presented below. In this example, the nonlinear model function is given by: $f(x,\theta) = 60 + 70e^{-x\theta}$ with the following experimental designs: x_1 = 4 and x_2 = 41. In Figure 4.4, the expectation surface is plotted for this design with marks for θ = 0.01, 0.02, .., 0.08, 0.1, 0.2, …, 0.9, 1.0. In Figure 4.5, the expectation curve with a different parameterization is plotted using ϕ = log θ, i.e. the nonlinear function is expressed by $f(x,\theta) = 60 + 70e^{-x10^{\phi}}$ with $\phi = -2.0, -1.9, ..., -0.1, 0$. First, the expectation surfaces in Figures 4.4 and 4.5 are identical and nonlinear regardless of the parametrization of the model (θ). This aspect reflects the intrinsic nonlinearity. On the other hand, the equally spaced values of θ did not translate to equally spaced points on the expectation curve. Nevertheless, this observation changes upon reparametrizing θ in terms of ϕ. Such parameterization dependent nonlinearity is called parameter-effects nonlinearity. Finally, Figure 4.6 presents the expectation curve for a different experimental design with x_1= 4 and x_2= 12 with the same θ values as in Figure 4.4. As it is seen, the different design affects the intrinsic nonlinearity, which is evident in the change of the shape of the curve. Intrinsic nonlinearity depends on the design and the parameter-effects nonlinearity depends on the parameterization.

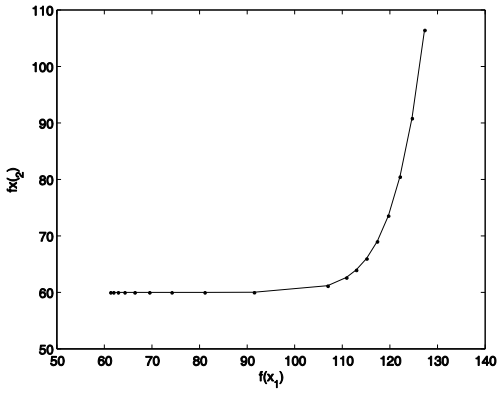

Figure 4.4. Expectation surface with x = (4, 41) and parameterization in terms of θ

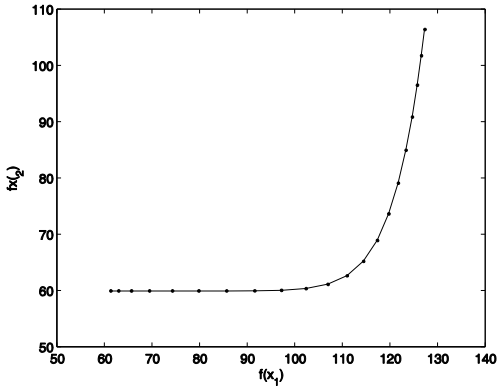

Figure 4.5. Expectation surface with x = (4, 41) and parameterization in terms of $\phi=\log_{10}\theta$

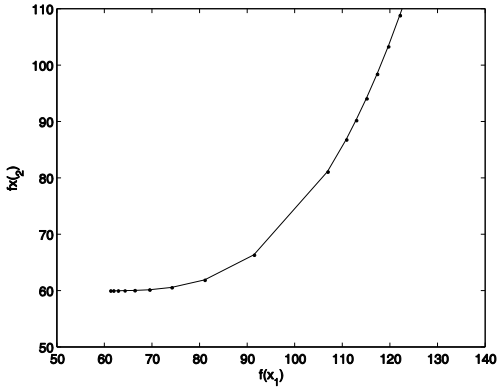

Figure 4.6. Expectation surface with x = (4, 12) and parameterization in terms of θ

To make the relative curvatures (Eq.(4.12) and (4.13)) scale-free, Bates and Watts [164] had used the scaling factor ρ, where $\rho = s\sqrt{p}$ and $s^2 = \left[\left(\mathbf{y}-\hat{\boldsymbol{\mu}}\right)^T\left(\mathbf{y}-\hat{\boldsymbol{\mu}}\right)\right]\Big/ n-p$. Following the same normalization procedure, the normalized relative curvatures are given by

$$\gamma_{\mathbf{h}}^{t} = K_{\mathbf{h}}^{\mathbf{t}}\rho, \quad \gamma_{\mathbf{h}}^{n} = K_{\mathbf{h}}^{n}\rho. \tag{4.16}$$

In addition, recasting **h** in the rotated axes as **h = Ld**, the tangent line $\dot{\boldsymbol{\mu}}_{\mathbf{Ld}}$ will have a unit norm (i.e. $\left\|\dot{\boldsymbol{\mu}}_{\mathbf{Ld}}\right\| = 1$) when **d** is a unit vector. In this case, the computation of $\gamma_{\mathbf{Ld}}^{t}$ and $\gamma_{\mathbf{Ld}}^{n}$ has been simplified into:

$$\gamma_{\mathbf{Ld}}^{t} = \rho\left\|\mathbf{d}^T\mathbf{A}_{\cdot\cdot}^{t}\mathbf{d}\right\|, \quad \gamma_{\mathbf{Ld}}^{n} = \rho\left\|\mathbf{d}^T\mathbf{A}_{\cdot\cdot}^{n}\mathbf{d}\right\| \quad \forall\mathbf{d}:\|\mathbf{d}\| = 1 \tag{4.17}$$

In the proposed experimental design, the maximum of these curvature measures are used, where

$$\gamma_{\max}^{t} = \max_{\|\mathbf{d}\|=1}\gamma_{\mathbf{Ld}}^{t} \tag{4.18}$$

and

$$\gamma_{\max}^{n} = \max_{\|\mathbf{d}\|=1}\gamma_{\mathbf{Ld}}^{n} \tag{4.19}$$

In formulating the MOO for the design of experiment, two general criteria have been taken into account. The first is that the experiment should be designed to maximize the informativeness of the data about model parameters. In this case, the standard D-optimal design is used as one of the objective functions, but other FIM-based metrics can also be used. As discussed above, the FIM-based designs will work well when the model outputs behave (somewhat) linearly with respect to the model parameters. To this end, the second

design criterion in the MOO aims to reduce model nonlinearity by minimizing the parameter-effects and intrinsic curvatures. The MOO formulation offers certain advantages over a combined that there is no prioritization of any one of the criteria and a Pareto optimal set of solutions can be obtained [169]. In this chapter, a new MBDOE is proposed using the following MOO problem. For the dynamical systems following Eq. (2.1), the MOO formulation of the design of experiments is given by:

$$\begin{aligned} &\max_{\mathbf{x}_0,\mathbf{t}_{sp},\mathbf{u}(t)} \prod_i \lambda_i \\ &\min_{\mathbf{x}_0,\mathbf{t}_{sp},\mathbf{u}(t)} \gamma^t_{\max} \\ &\min_{\mathbf{x}_0,\mathbf{t}_{sp},\mathbf{u}(t)} \gamma^n_{\max} \end{aligned} \tag{4.20}$$

subject to

$$\begin{aligned} &\frac{d\mathbf{x}\left(t,\hat{\boldsymbol{\theta}}\right)}{dt} = \mathbf{g}\left(\mathbf{x}\left(t,\hat{\boldsymbol{\theta}}\right),\mathbf{u},\hat{\boldsymbol{\theta}}\right) \\ &\mathbf{x}\big|_{t=0} = \mathbf{x}_0 \\ &\mathbf{x}_0^L \le \mathbf{x}_0 \le \mathbf{x}_0^U \\ &\mathbf{u}_j^L \le \mathbf{u}_j \le \mathbf{u}_j^U \end{aligned} \tag{4.20}$$

where λ_i is the i-th eigenvalue of the $\text{FIM} = \hat{\mathbf{F}}_{\bullet}^T \mathbf{V}^{-1} \hat{\mathbf{F}}_{\bullet}$. The parameter vector $\hat{\boldsymbol{\theta}}$ is either an initial guess of the parameter values or the parameter estimates from the current iteration in an iterative model identification procedure[4]. The decision variables include the initial condition of the states ($\mathbf{x}_0$), the sampling time points of measurements ($\mathbf{t}_{sp}$), and the piecewise-constant dynamic input $\mathbf{u}(t)$. In the case studies below, the input $\mathbf{u}(t)$ was parameterized using a zero-order hold, defined by the vectors of magnitudes ($\mathbf{u}_j$, j = 1, 2,…, m) and switching times ($\mathbf{t}_{sw,j}$) (see Figure 4.1). The number of measurement time points and that of input switching times were prescribed prior to the design. Note that the

minimization of other measures of curvatures, such as the root mean square (RMS) of the Hessian coefficients proposed by Benabbas *et al.* [158], can be also used in place of the last two objective functions.

The proposed multi-objective design criterion is compared with the Q-optimality. Hamilton and Watts [165] proposed the Q-optimality by minimizing the volume of the second-order approximation of parameter inference region, as given below:

$$v\left(\varphi,\hat{\boldsymbol{\theta}}\right)=c\left|\hat{\mathbf{F}}_{\cdot}'\hat{\mathbf{F}}_{\cdot}\right|^{-1/2}\left|\mathbf{D}\right|^{-1/2}\left\{1+k^{2}\mathrm{tr}\left(\mathbf{D}^{-1}\mathbf{M}\right)\right\}. \tag{4.21}$$

where the matrices **D** and **M** involve the intrinsic and parameter-effects curvatures, respectively, and are functions of $\hat{\mathbf{F}}_{\cdot\cdot}$, $c=\dfrac{\pi^{p/2}\rho_{\alpha}}{\Gamma\left[0.5*\left(p+2\right)\right]}$, $k=\rho_{\alpha}\Big/\sqrt{2\left(p+2\right)}$. The parameter k is the effective noise level which denotes the 100(1-α)% point of χ^{2} distribution with *p* degrees of freedom. This second order approximation is the product of a linear approximation $\left|\hat{\mathbf{F}}_{\cdot}'\hat{\mathbf{F}}_{\cdot}\right|^{-1/2}$ and a quadratic polynomial whose coefficients depend on the parameter-effects curvature and intrinsic curvature.

4.4.2 Numerical Implementation of MOO

As described in the previous section, the intrinsic and parameter-effects curvatures require the computation of the first and second-order model sensitivities. For the ODE model in Eq. (3), the first-order sensitivities can be calculated according to:

$$\hat{\mathbf{F}}_{\cdot}=\mathbf{F}_{\cdot}\left(\hat{\boldsymbol{\theta}};\mathbf{x}\right)=\frac{\partial\mathbf{F}\left(\mathbf{x}\left(t,\mathbf{u},\boldsymbol{\theta}\right)\right)}{\partial\mathbf{x}}\frac{\partial\mathbf{x}\left(t,\mathbf{u},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}/\boldsymbol{\theta}}\Bigg|_{\hat{\boldsymbol{\theta}}} \tag{4.21}$$

Note that the sensitivities in the above equation are normalized with respect to the parameter values. The last term on the right hand side is the first-order sensitivities of the ODE model, which obey the following differential equation:

$$\begin{aligned} &\frac{d}{dt}\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{g}}{\partial \mathbf{x}}\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\theta}} \\ &\left.\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}\right|_{t=0} = 0 \end{aligned} \tag{4.21}$$

Here, we have assumed that $\mathbf{x}_0$ is known, but such assumption can be easily relaxed in the case that the initial conditions are to be estimated from data (i.e. the initial conditions are treated as unknown parameters). In this study, the sensitivities $\partial \mathbf{x}/\partial \boldsymbol{\theta}$ were computed by solving the ODE in Eq. (4.21) simultaneously with that in Eq. (2.1), following a procedure known as the direct differential method (DDM) [122]. Meanwhile, the Hessian matrix was approximated using a finite difference method, as follows:

$$\hat{\mathbf{F}}_{\bullet\bullet ijk} \approx \begin{cases} \dfrac{F_i\left(\boldsymbol{\theta}+\Delta\theta_j\mathbf{e}_j\right)-2F_i\left(\boldsymbol{\theta}\right)+F_i\left(\boldsymbol{\theta}-\Delta\theta_j\mathbf{e}_j\right)}{\Delta\theta_j^2/\theta_j^2} & \text{for } j=k \\ \dfrac{F_i\left(\boldsymbol{\theta}+\Delta\theta_j\mathbf{e}_j+\Delta\theta_k\mathbf{e}_k\right)-F_i\left(\boldsymbol{\theta}+\Delta\theta_j\mathbf{e}_j-\Delta\theta_k\mathbf{e}_k\right)-F_i\left(\boldsymbol{\theta}-\Delta\theta_j\mathbf{e}_j+\Delta\theta_k\mathbf{e}_k\right)+F_i\left(\boldsymbol{\theta}-\Delta\theta_j\mathbf{e}_j-\Delta\theta_k\mathbf{e}_k\right)}{\left(\Delta\theta_j/\theta_j\right)\left(\Delta\theta_k/\theta_k\right)} & \text{for } j\neq k \end{cases} \tag{4.21}$$

where $\mathbf{e}_j$ is the j-th elementary vector and using 1% parameter perturbations (i.e. $\Delta\theta_j/\theta_j = 0.01$). The second order sensitivities above are again normalized with respect to the model parameters for the same reason as in Eq. (4.21).

In the case studies below, the MOO problem was solved using the genetic algorithm subroutine (*gamultiobj*) within the Optimization toolbox in MATLAB. The maximization of the determinant of FIM in the MOO was converted to the minimization of its negative. The optimal experimental design was selected from the Pareto front by balancing the

trade-offs among the objective functions. Briefly, for each member of the Pareto set, a ranking score was assigned based on the distances of the objective function values from their respective overall minimum. Specifically, the distance metric was computed as the difference between each objective function value and the minimum of the respective objective function over the entire Pareto set, normalized such that the distance metric lies between 0 and 1. The ranking score was subsequently evaluated as the root mean square of the distance metrics for all three objective functions, and the chosen Pareto design was one with the lowest ranking score. Of course, other selection strategy could be applied, such as the rough set method (RSM)[170] and different selection criteria may work equally well.

4.5 RESULTS AND DISCUSSION

The performance of the new MBDOE was evaluated through applications to two examples: a biodiesel production process [171] and a fed-batch fermentation of Baker's yeast [158], and compared with three other MBDOE designs: D-optimal, Q-optimal [165], and D-optimal with a constraint on the RMS of the Hessian coefficients [158].

4.5.1 CASE STUDY 1: BIODIESEL PRODUCTION PROCESS

The first case study was taken from the kinetic modeling of alkaline transesterification of vegetable oils into biodiesel [172]. Biodiesel consists of a mixture of 6 to 7 mono-alky methyl-esters, which derive from fatty acids with long chains of carbon atoms. Transesterification, also known as alcoholysis, is commonly used to convert triglycerides into biodiesel, where in the presence of a catalyst, such as sodium

methoxide, methanol is used to chemically break the oil molecules into methyl esters and byproduct glycerol. The triglycerides (TG) are thus converted stepwise into diglycerides (DG), monoglycerides (MG) and finally to glycerol, and a mole of ester is produced in each step as shown in the following reaction scheme.

$$\begin{aligned} &TG + MeOH \xleftrightarrow{k_1,k_2} DG + E \\ &DG + MeOH \xleftrightarrow{k_3,k_4} MG + E \\ &MG + MeOH \xleftrightarrow{k_5,k_6} GL + E \end{aligned} \tag{4.22}$$

The model equations used for this reaction scheme is given below

$$\begin{aligned} \frac{dM_1}{dt} &= -k_1 \frac{M_1 M_2}{\text{Vol}} + k_2 \frac{M_3 M_6}{\text{Vol}} \\ \frac{dM_2}{dt} &= u - k_1 \frac{M_1 M_2}{\text{Vol}} + k_2 \frac{M_3 M_6}{\text{Vol}} - k_3 \frac{M_3 M_2}{\text{Vol}} + k_4 \frac{M_4 M_6}{\text{Vol}} - k_5 \frac{M_4 M_2}{\text{Vol}} + k_6 \frac{M_5 M_6}{\text{Vol}} \\ \frac{dM_3}{dt} &= k_1 \frac{M_1 M_2}{\text{Vol}} - k_2 \frac{M_3 M_6}{\text{Vol}} - k_3 \frac{M_3 M_2}{\text{Vol}} + k_4 \frac{M_4 M_6}{\text{Vol}} \\ \frac{dM_4}{dt} &= k_3 \frac{M_3 M_2}{\text{Vol}} - k_4 \frac{M_4 M_6}{\text{Vol}} - k_5 \frac{M_4 M_2}{\text{Vol}} + k_6 \frac{M_5 M_6}{\text{Vol}} \\ \frac{dM_5}{dt} &= k_5 \frac{M_4 M_2}{\text{Vol}} - k_6 \frac{M_5 M_6}{\text{Vol}} \\ \frac{dM_6}{dt} &= k_1 \frac{M_1 M_2}{\text{Vol}} - k_2 \frac{M_3 M_6}{\text{Vol}} + k_3 \frac{M_3 M_2}{\text{Vol}} - k_4 \frac{M_4 M_6}{\text{Vol}} + k_5 \frac{M_4 M_2}{\text{Vol}} - k_6 \frac{M_5 M_6}{\text{Vol}} \\ \text{Vol} &= \sum_{i=1}^{6} \frac{M_i}{\rho}. \end{aligned} \tag{4.23}$$

where the states M_i refer to triglycerides (M_1), methanol (M_2), diglycerides (M_3), monoglycerides (M_4), glycerol (M_5) and ester (M_6), u is the methanol input flow rate, Vol is the volume of the reaction mixture, k_j is the kinetic constant of the j-th reaction, and ρ_i is the density of the component M_i. The MBDOE were done for a 120-minute fed-batch experiment using the four different criteria mentioned above. Table 4.2 lists the model

parameters, the initial condition of the reactor, and the DOE design variables with their bounds, following those used by Franceschini and Macchietto [171].

Table 4.2. Initial guess and range for the design variables of the Case study 1

Variable	Symbol	Initial Guess	Lower bound	Upper bound
Initial amount of methanol (mol)	M_2^0	0.5656	0.5656	3.4
Sampling times (min)	$\mathbf{t}_{sp}$	[5, 30, 60, 80, 90,. 100, 100]	3	120
Switching time (min)	$\mathbf{t}_{sw}$	[20, 40, 65, 90]	0.1	120
Flow rate of methanol (mol/min)	**u**	[0.0273, 0.0273, 0.0273, 0.0273, 0]	0	2

4.5.2 Case Study 2: Baker's Fermentation of Yeast

The second application was taken from a model of a fed-batch fermenter [173]. In the model, cellular growth and product formation, as described by biomass x_1, are assumed to rely on a single substrate x_2. Furthermore, the fermenter is assumed to operate at a constant temperature and the feed is free from product. Assuming Monod-type growth kinetics, the model equations are given below.

$$\frac{dx_1}{dt} = (r - u_1 - \theta_4)x_1$$
$$\frac{dx_2}{dt} = -\frac{rx_1}{\theta_3} + u_1(u_2 - x_2) \qquad (4.24)$$
$$r = \frac{\theta_1 x_2}{\theta_2 + x_2}$$
$$x_1 = \text{biomass};\ x_2 = \text{substrate}$$

where x_1 is the biomass concentration (g/L), x_2 is the substrate concentration (g/L), u_1 is the dilution factor (h^{-1}) and u_2 is the substrate concentration in the feed (g/L). The parameter values, initial conditions and constraints for the design variables can be found in Table 4.3, following the values from a previous publication [158].

Table 4.3. Range for the design variables of the Case study 2

Variable	Symbol	Lower bound	Upper bound
Biomass initial condition	x_1^0	1	10
Sampling times (h) (max 20 intervals)	$\mathbf{t}_{sp}$	0	40
Dilution factor switching time (h) (max 10 switching intervals)	$\mathbf{t}_{sw1}$	0	40
Feed substrate switching time (h) (max 10 switching intervals)	$\mathbf{t}_{sw2}$	0	40
Dilution factor control levels (h^{-1})	$\mathbf{u}_1$	0.05	0.2
Feed substrate control levels (g/L)	$\mathbf{u}_2$	5	35

4.5.3 PERFORMANCE EVALUATION

The performance of each criterion was evaluated using the number of parameters that were *a priori* and practically identifiable from the respective optimal experiment. As mentioned in Chapter 2, there are two types of parameter identifiability; the first assumes noise-free (ideal) data, referred to as *a priori* identifiability, while the other accounts for random noise in data, referred to as practical identifiability. The methods used for assessing each type of parameter identifiability that were presented in Chapter 2 were used in this work. Figures 4.7-4.9 show the optimal input profiles for the two case studies and Tables 4.4 and 4.5 summarize the performance of each criterion in terms of the number of identifiable parameters.

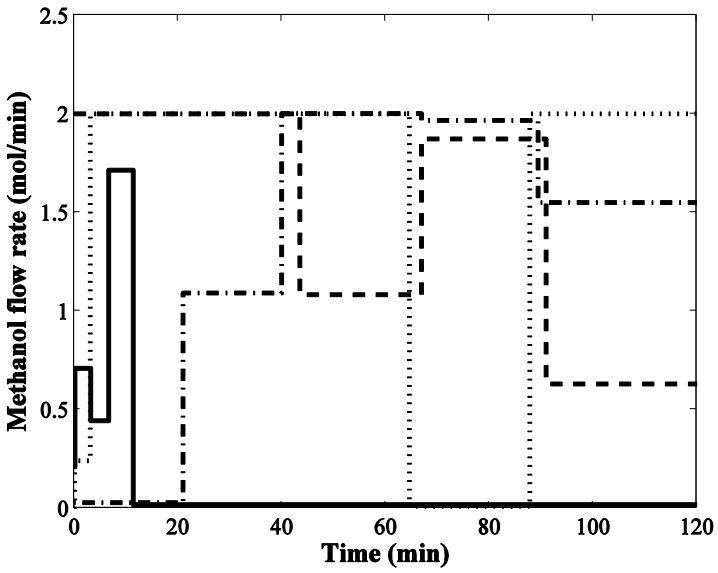

Figure 4.7. Optimal methanol flow rates (*u*) in Case Study 1 from the model based design of experiments. (- - -) D-optimality, (- · -) Q-optimality, (···) represents Threshold curvature and (—) MOO design.

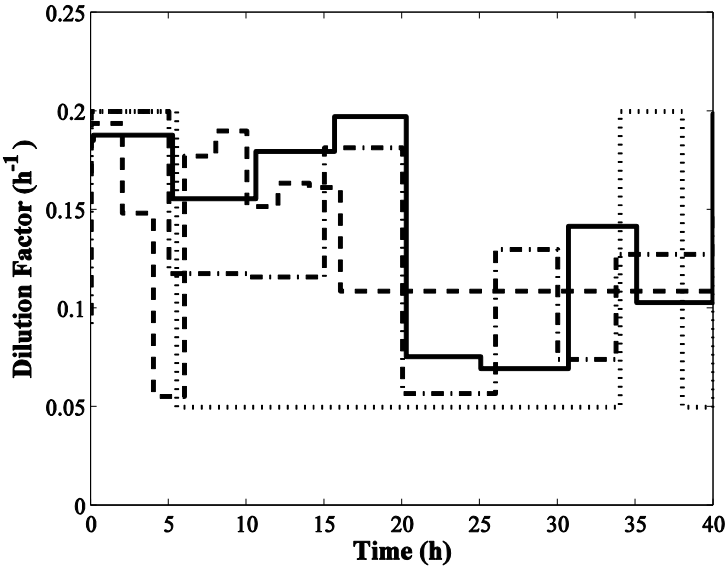

Figure 4.8. Optimal dilution factor (u_1) in Case Study 2 from the model based design of experiments. (- - -) D-optimality, (- · -) Q-optimality, (···) represents Threshold curvature and (—) MOO design.

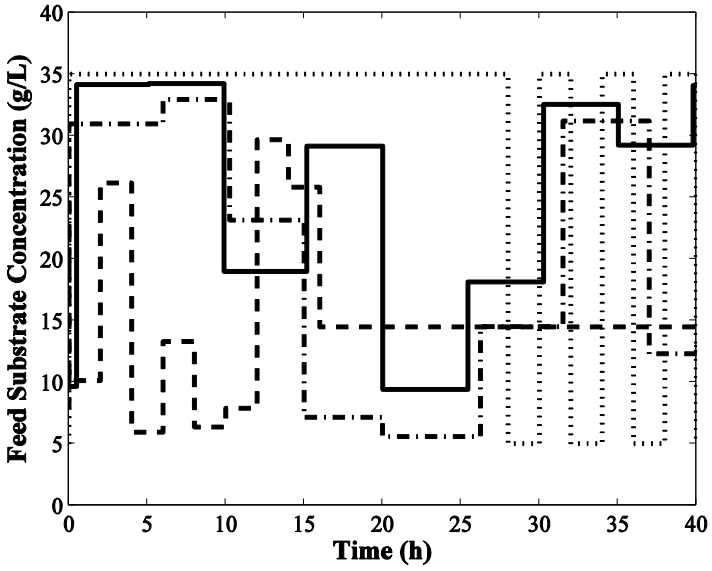

Figure 4.9. Optimal feed substrate concentration (u_2) in Case Study 2 from the model based design of experiments. (- - -) D-optimality, (- · -) Q-optimality, (···) represents Threshold curvature and (—) MOO design.

In general, the MOO and Q-optimal designs gave the highest number of *a priori* and practically identifiable parameters, with the MOO having a slight advantage over the Q-optimal, while D-optimal experiments were consistently the poorest performing design among the four methods compared. The results clearly demonstrated the advantage of reducing the effect of parametric and intrinsic curvature in the design of experiment that relies on model linearization. At the same time, the D-optimal design was clearly suboptimal in both case studies, especially when data noise is important (see practical identifiable parameters), as model nonlinearity was not accounted for. In the second case study, one should, at least in theory, be able to estimate all parameters from ideal and noisy data using the conditions designed by the MOO criterion. Nevertheless, the main drawback of any curvature-based methods is the higher computational requirement to compute the Hessian of a dynamical model than any of the FIM-based designs. However, this disadvantage is not prohibitive as the design of experiment is typically done offline.

Table 4.4. Number of Identifiable Parameters in Case Study 1 (Total parameters = 6)

Design	AIP*	PIP[#]
D-optimality	3/6	1/6
Threshold curvature	4/6	2/6
Q-optimality	4/6	4/6
Multi-Objective	4/6	5/6

*AIP: *A priori* Identifiable Parameters

[#]PIP: Practically Identifiable Parameters

Table 4.5. Number of Identifiable Parameters in Case Study 2 (Total parameters = 4)

Design	AIP	PIP
D-optimality	3/4	2/4
Threshold curvature	2/4	3/4
Q-optimality	3/4	4/4
Multi-objective	4/4	4/4

Table 4.6. Parameter estimates calculated from all the four designs for Case Study 2

True Parameters	D optimality	Threshold Curvature	Q-optimality	MOO
0.3	0.488	0.4320	0.3712	0.3039
0.2	0.6328	0.8560	0.9273	0.2062
0.5	0.6646	0.5273	0.4444	0.5070
0.05	0.0955	0.0733	0.0545	0.0511

As a proof of principle, a parameter estimation of the Baker's yeast fermentation model (second case study) was carried out by simulating the model with the true parameters in Table 4.6 and according to the experimental designs in Figures 4.8-4.9. The simulated data were contaminated with 30% i.i.d. Gaussian noise. The least square error minimization problem was solved using the constrained optimization subroutine *fmincon* in MATLAB with the true parameter values as the initial guess. In this case, the accuracy of the parameter estimates from different methods should depend on the informativeness

of the experimental designs and not on the effectiveness of the optimization algorithm. Table 4.6 summarizes the result of the parameter estimation, which reflects the parameter identifiability comparison above. The curvature-based methods have comparable performance, while the MOO design gave the most informative data for estimating the model parameters. The Pareto surface and the state trajectory associated with the MOO design in each case study are shown in Figures 4.10 and 4.11, respectively.

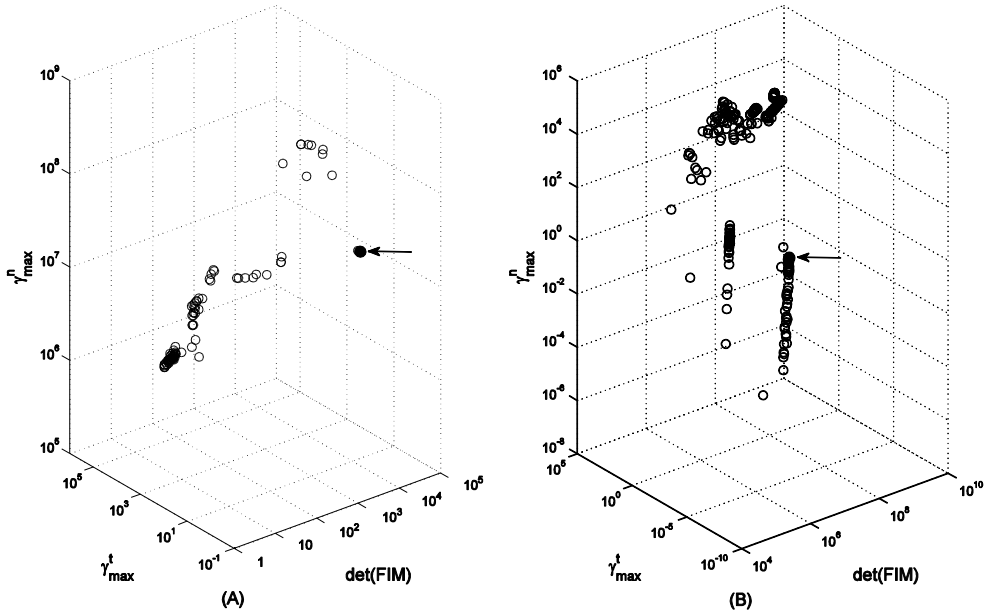

Figure 4.10. Pareto fronts for (A) Case Study 1 and (B) Case Study 2. The chosen optimal Pareto designs for the case studies are indicated by the arrows (filled circles).

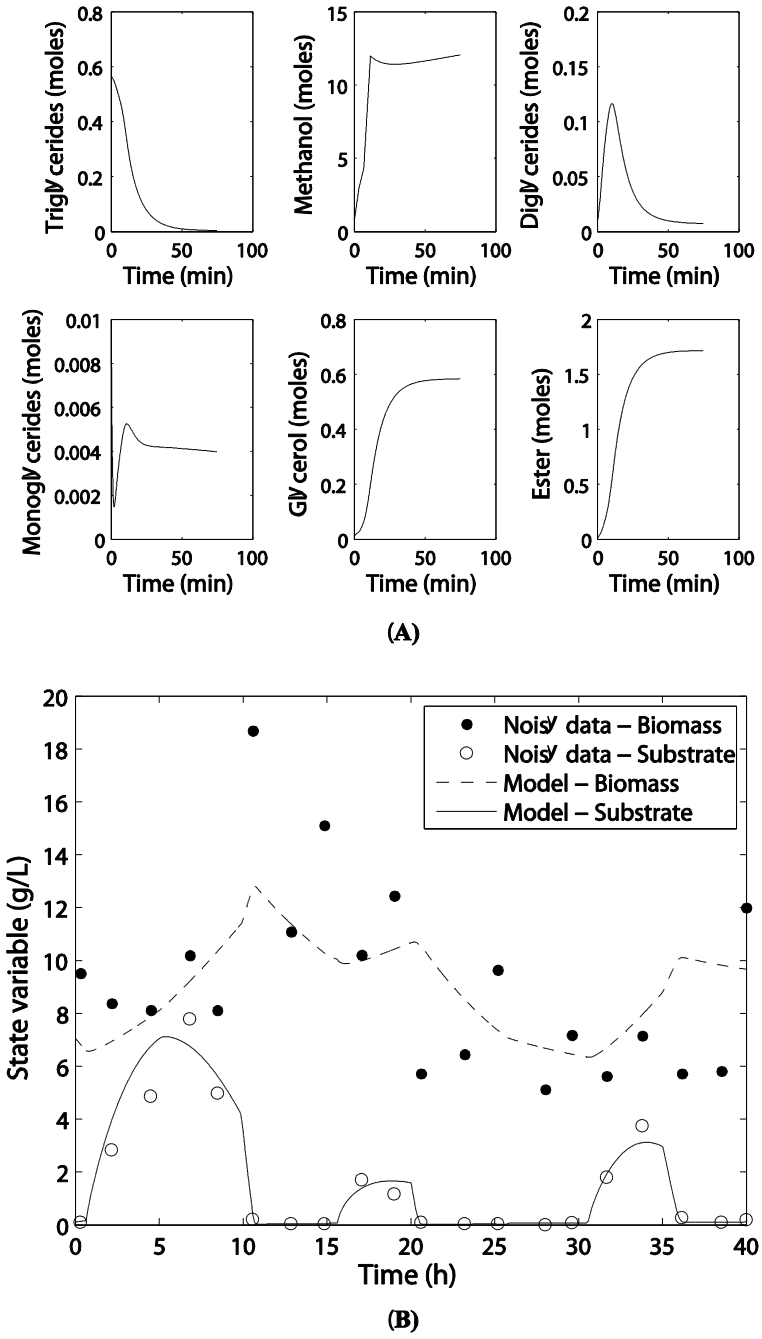

Figure 4.11. State trajectories of the system in (A) Case Study 1 and (B) Case Study 2 under the optimal experiment conditions from the MOO design. The parameter estimation for Case Study 2 was performed using computer-generated noisy data, as shown in (B).

4.6 CONCLUSIONS

A new model-based design of experiment was formulated, which takes into account model nonlinearity through the parametric and intrinsic curvature effects. The curvature measures were based on the second-order sensitivity (Hessian) of the model with respect to its parameters. The proposed MBDOE relied on the simultaneous minimization of model curvature and maximization of data informativeness using a multi objective optimization framework. Applications to two case studies of dynamical models demonstrated that the MOO criterion outperformed FIM-based D-optimal design as well as other curvature-based designs, including Q-optimality, in terms of the number of identifiable parameters from each respective optimal experiment. The proposed MOO formulation is flexible and can be easily modified to accommodate other design criteria that may become important in certain applications.

CHAPTER 5

ITERATIVE MODEL IDENTIFICATION

5.1 INTRODUCTION

The final contribution of this thesis is in the integration of identifiability analysis, parameter estimation and design of experiments into a complete model identification tool. Given the data and preliminary model equations, this tool will be able to automate the iterative procedure of performing optimal experiment design, applying identifiability analysis and finally getting better parameter estimates. While the concept of iterative model identification is well established, the novelty comes in the way these different components are integrated and the methods developed for identifiability analyses and model based design of experiments.

Iterative process for model identification has been proposed by many researchers in the field of systems biology and was also highlighted by Banga and co-workers in a recent article [88]. Rabitz and co-workers proposed an iterative model identification process [82] wherein they used a closed-loop strategy to estimate how to stimulate a process and how to observe the system for identification purpose. Gadkar *et al* [131] proposed an alternative identification procedure involving selection of species whose concentration measurements would aid the model calibration and model discrimination. In reality, there are often additional practical constraints, for example not all components of the system can be measured and only specific stimuli are available, which makes the parameter identification problem even severe. These constraints along with the typical

dynamic nonlinear behavior of the system cause the parameter identifiability issue. The proposed strategy, described in the following section, has a strong emphasis on improving the parameter identifiability iteratively. This strategy basically has three steps (1) performing model-based design of experiments with a random parameter values as the starting point and then generating data using the obtained design and the true parameter values (2) carrying out prior identifiability analysis using the optimal experimental design and (3) estimating the parameters using a two-phase parameter estimation strategy guided by a post-estimation identifiability analysis. This iterative model identification strategy has been applied to a five variable gene regulatory network [66], which is modeled using the S-systems, to illustrate intricacies in each step of the model identification cycle and the methodology of integrating these steps together.

5.2 METHODS

Usually, any model development process is started off with an objective or problem definition. The biological model identification cycle is then proceeded with the gathering of background knowledge about the biological system and data of the pathway/network of interest from the literature. The next step is to choose a model structure and to write model equations that simulate the biochemical pathway/network of interest. In this case, ODEs are the most commonly used modeling equations, which are typically written to describe balance equations (e.g. mass or mole balance). In this chapter, model identification is started from scratch by assuming no prior knowledge of parameter values. So, a set of guess values are taken from a uniform random distribution and used in the first iteration of the iterative procedure as shown in Figure 5.1. This

figure is quite similar to Figure 1.2 and as mentioned in that section the individual processes in the box and their integration are of particular focus in this chapter

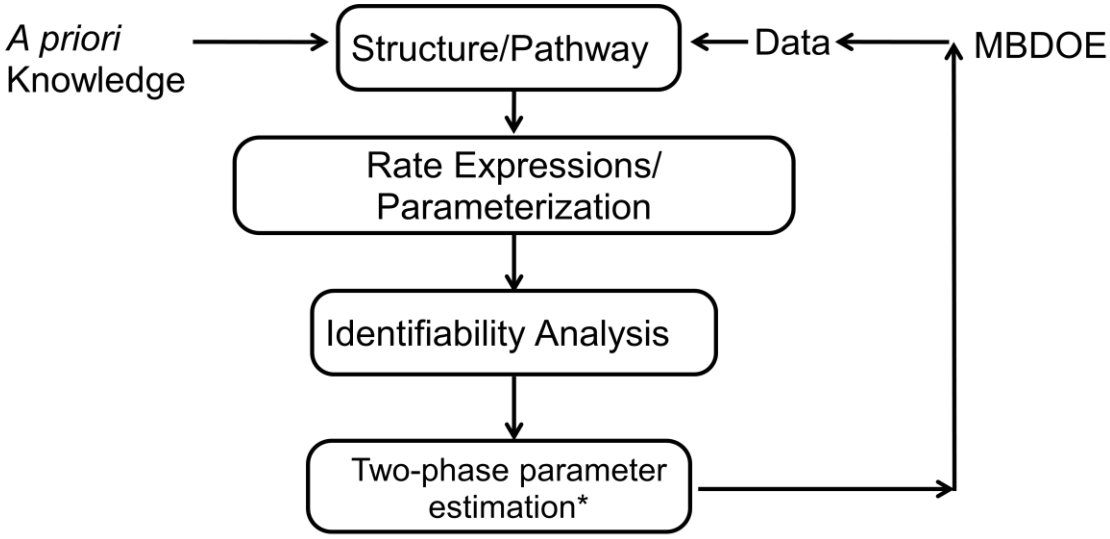

Figure 5.1. The iterative model development procedure adopted in this work

The first key step in the procedure is to perform MBDOE using the guessed parameters and the methods described in Chapter 4. Three MBDOEs are compared in this chapter, including the D-optimality, Q-optimality or MOO designs. The obtained design is used to generate data for parameter estimation. The next key step is to carry out parameter identifiability analyses. As discussed in Chapter 2, there are two main reasons to carry out this identifiability analysis before estimating the parameters from the data. Firstly, these parameters have some biological meaning and hence there arises a necessity to know whether the parameter estimates are reliable in order to avoid wrong inferences made based on these parameters. Secondly, parameters that are unidentifiable in theory cannot be estimated, and this problem can manifest in severe issues in the numerical optimization. It is worth mentioning that in this iterative procedure *a priori* identifiability

analysis is carried out using the method described in Section 2.1.5.7 and the practical identifiability analysis was carried using the Method 2 described in Section 2.2.1.2

The subsequent step is to estimate the parameter values based on the information obtained in the *a priori* identifiability analysis. In this chapter, a two-phase parameter estimation strategy [39] was adopted. Although the two-phase strategy was developed to tackle missing data problem, the method is used for a different purpose here. The first phase of this estimation employs the decoupled ODE parameter estimation, as discussed in Chapter 3, to obtain the subset of parameters that have been identified as non-AIP. These parameters are then fixed in the second phase in which the remaining parameters (i.e. those that are identified as AIP) are estimated by the ODE decomposition method [174]. In the ODE decomposition estimation, the ODE model is solved one equation at a time, and similar to decoupling method, this method also decouples the parameter estimation problem. The main reason for estimating only the AIP in the second phase is not only that these are the maximal set of parameters that can be uniquely identified by definition, but also to reduce the dimensionality of the parameter search space, thereby allowing a faster convergence of the search algorithm to the globally optimal solution. With the new parameter estimates, the whole procedure is repeated until convergence.

5.3 CASE STUDY

The case-study considered below represents an ideal system which captures the essential features of many actual biological systems. The model is taken from an article by Hlavacek and Savageau [175], which describes a gene regulatory network with two genes: a regulator gene and an effector gene. The dynamics in the concentrations arises

from the processes of transcription, translation, specific degradation, dilution and metabolism. These processes are depicted in Figure 5.2. The network is modeled using five components: the concentrations of effector gene mRNA (X_1), enzyme (X_2), inducer (X_3), regulator gene mRNA (X_4) and regulator protein (X_5). The precursor pools for mRNA and protein biosynthesis, X_6 and X_7 are assumed to be maintained at constant levels throughout the experiment. The substrate concentration X_8 is also maintained constant. This system was modeled using the S-systems, given by: [176]

$$\begin{aligned}
\dot{\mathbf{X}}_1 &= \alpha_1 X_3^{g_{13}} X_5^{g_{15}} X_6^{g_{16}} - \beta_1 X_1^{h_{11}} \\
\dot{\mathbf{X}}_2 &= \alpha_2 X_1^{g_{21}} X_7^{g_{27}} - \beta_2 X_2^{h_{22}} \\
\dot{\mathbf{X}}_3 &= \alpha_3 X_2^{g_{32}} X_8^{g_{38}} - \beta_3 X_2^{h_{32}} X_3^{h_{33}} \\
\dot{\mathbf{X}}_4 &= \alpha_4 X_3^{g_{43}} X_5^{g_{45}} X_6^{g_{46}} - \beta_4 X_4^{h_{44}} \\
\dot{\mathbf{X}}_5 &= \alpha_5 X_4^{g_{54}} X_7^{g_{47}} - \beta_5 X_5^{h_{55}}
\end{aligned} \tag{5.1}$$

where X_i denotes the concentration of the *i*-th component shown in Figure 5.2. The initial conditions are [0.7 0.12 0.14 0.16 0.18]. The quantities from X_6-X_8 were considered as independent variables and set to 1 in the original model. The true parameter values for the purpose of generating *in silico* data in this exercise are given below in the Table 5.1.

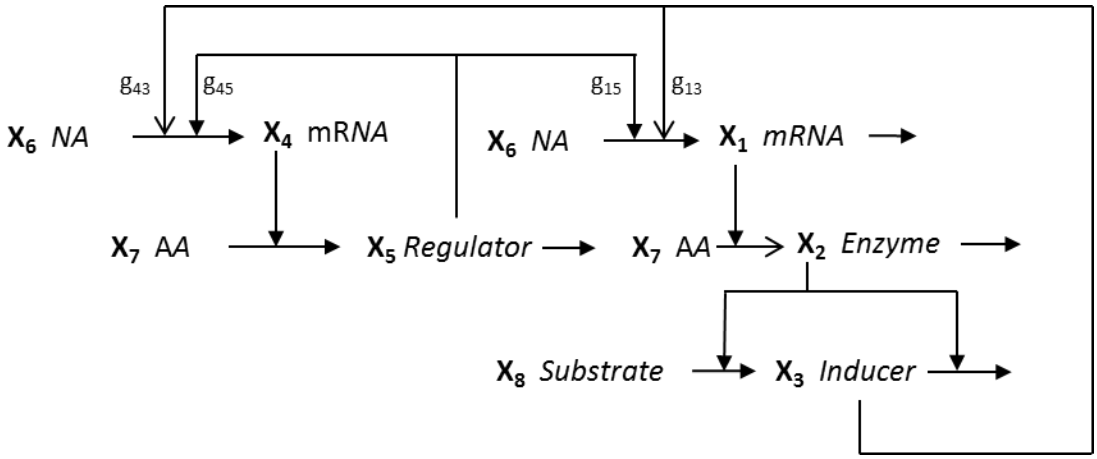

Figure 5.2. Ideal regulatory system. The horizontal arrows indicate the direction of transcription. The vertical arrows indicate the influences of a regulator (filled arrow heads) and an inducer (normal arrow heads).

Table 5.1. True parameter values of the gene regulatory pathway

i	α_i	g_{i1}	g_{i2}	g_{i3}	g_{i4}	g_{i5}	β_i	h_{i1}	h_{i2}	h_{i3}	h_{i4}	h_{i5}
1	5	0	0	1	0	-1	10	2	0	0	0	0
2	10	2	0	0	0	0	10	0	2	0	0	0
3	10	0	-1	0	0	0	10	0	-1	2	0	0
4	8	0	0	2	0	-1	10	0	0	0	2	0
5	10	0	0	0	2	0	10	0	0	0	0	2

5.3.1 Iterative Model Identification

The following constraints are imposed in the optimal experiment design

- No more than five stimulus steps within the bounds $1 \leq X_8 \leq 3$ of the substrate concentration per experiment,

- no more than 10 sampling points per experiment,
- all the dependent variables' concentrations are available (no missing data),

The iterative model identification is initiated with a design of experiment step using a parameter set randomly sampled from uniform distribution between 0 to 20 for rate constants and -2 to 2 for kinetic orders. This will be the initial guess (IG) for the parameters. Three sets of IGs are used to test the reproducibility of the method. After obtaining the optimal design in the first step, data is then generated by simulating the model with this design and the true parameter values. The generated data are then contaminated with 5% i.i.d. Gaussian noise. The noisy data are then used for parameter estimation, guided by identifiability analyses, as described above. The series of steps is repeated. For this example, it is found that the set of identifiable parameters and parameter estimates converge after 6 iterations.

Table 5.2. Summary of identifiability results at the end of iterative model identification for IG1

Design	AIP	PIP
D-optimality	15/23	11/15
Q-optimality	16/23	14/16
MOO	18/23	16/18

The number of iterations (N_{exp}) was finalized to be six because most of parameters seemed to converge to the true value after the sixth experiment. However, deciding the number of iterations for a new problem could be tricky. Approaches such as those proposed by Marino and Voit [48], wherein the authors come up with a framework for automated procedure for the extraction of metabolic information from time-series data for

BST modeling framework, can be adopted to tackle this problem. The summary of identifiability results for all three designs is given in Table 5.2. As it can be seen from Table 5.2, MOO design clearly outperforms the other two design criteria in terms of parameter identifiability. Using MOO, after 6 iterations, 21 out of the 23 parameters were *a priori* identifiable, and all of these 21 AIPs were practically identifiable. This points to the fact that the parameter estimates yielded by the MOO design are less uncertain than the parameters obtained using the other two designs. The set of AIPs at the end of each iteration are summarized in Table 5.3 for each of design criteria considered. As it can be seen, D-optimal designs led to the largest of parameter errors, while the MOO design was able to provide parameter estimates with the smallest mean and standard deviation of the relative error (reported in Tables 5.4-5.6). It is evident from these results that irrespective of the design criterion used, the iterative model identification is able to (gradually) increase the number of identifiable parameters. However, methods that account for the curvature (using sensitivity and Hessian matrix) performed better than FIM-based method (using only sensitivity matrix). The relative parameter errors and the number of identifiable parameters were progressively decreasing and increasing, respectively, with the number of iteration (see Tables 5.3-5.6). The parameter estimates at the end of each iteration for all the three designs are reported in Tables 5.10-5.12. The performance of the MOO is also reproducible upon restarting of the iterations using different random initial guess values for the parameters. The summary of identifiability results, AIPs and relative parameter errors after each iteration for two other initial guesses, IG2 and IG3, are presented in Appendix B.

Table 5.3. No. of AIPs after each iteration for all the designs for IG1.

Design	Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
D-optimality	6	6	10	14	14	15
Q-optimality	6	13	13	14	15	16
MOO	7	9	13	17	18	18

Table 5.4. Mean of the relative errors of the parameter estimates of only AIP for IG1

Design	Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
D-optimality	57.84	41.41	33.51	19.35	11.59	4.54
Q-optimality	22.34	45.34	28.31	10.83	2.13	1.81
MOO	46.81	35.92	22.26	6.31	2.20	1.02

Table 5.5. Standard deviation of the relative errors of the parameter estimates of only AIP for IG1

Design	Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
D-optimality	34.64	31.09	28.41	26.53	25.75	4.41
Q-optimality	23.29	29.22	27.46	9.33	1.81	1.96
MOO	34.83	30.46	18.90	5.34	3.67	0.85

Table 5.6. Maximum of the relative errors of the parameter estimates of only AIP for IG1

Design	Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
D-optimality	94.59	89.4	98.22	99.89	100	13.14
Q-optimality	59.62	108.9	98.33	31.14	6.48	7.33
MOO	95.49	89.76	64.50	17.87	16.3	2.5

Table 5.7. AIP after each iteration for D-optimality design and IG1

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
2	2	2	1	1	1
14	5	5	2	2	2
15	14	12	4	4	4
20	15	13	5	5	5
21	21	14	7	7	7
23	23	15	12	12	11
		16	13	13	12
		21	14	14	13
		22	15	16	14
		23	16	19	15
			20	20	16
			21	21	19
			22	22	21
			23	23	22
					23

Table 5.8. AIP after each iteration for Q-optimality design and IG1

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
2	3	2	2	2	2
5	4	3	3	3	3
6	5	5	4	4	4
8	6	6	5	5	5
12	7	9	6	6	6
14	11	11	11	7	7
	12	12	12	11	10
	13	13	13	12	11
	16	16	14	13	12
	18	18	16	14	13
	19	19	18	16	14
	20	20	19	18	16
	21	21	20	19	18
			21	20	19
				21	20
					21

Table 5.9. AIP after each iteration for MOO design and IG1

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
2	1	1	1	1	1
5	2	2	2	2	2
6	5	4	4	4	4
12	12	5	5	5	5
14	14	6	6	6	6
15	15	12	8	7	7
21	19	14	11	8	8
	21	15	12	11	11
	22	16	13	12	12
		17	14	13	13
		19	15	14	14
		20	16	15	15
		21	17	16	16
			19	17	17
			20	19	19
			21	20	20
			22	21	21
				22	22

Table 5.10. Parameter estimates after each iteration of D-optimality design for IG1

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
2.0662	2.8862	4.2037	4.4681	4.7112	4.7113
8.7497	8.9816	9.4573	9.4342	9.6935	9.7500
12.1807	9.1284	0.1201	2.3242	2.4368	4.8567
16.5554	20.0000	7.1549	7.5748	8.0834	8.0882
20.0000	14.3040	11.9652	10.9863	10.1994	10.1996
8.1866	0.0037	1.8133	4.1085	0.7656	11.1406
2.1031	2.5040	6.6671	8.5610	8.6839	8.6865
1.5427	0.6287	1.2368	6.2336	14.7664	19.0181
4.4971	1.3643	7.5793	19.9957	13.9782	19.7471
13.5806	20.0000	20.0000	20.0000	19.9999	20.0000
0.0153	0.2624	0.1361	0.7315	0.8422	0.8726
-1.5614	-1.5405	-0.6472	-1.1770	-1.0051	-0.9978
1.4563	2.0982	2.1916	1.7389	1.8996	1.8986
-0.0541	-0.1060	-0.4567	-0.9997	-1.0666	-0.9999
0.3482	1.3500	1.4644	1.6115	2.0800	2.0800
-0.5160	-1.6806	-1.4600	-0.9074	-0.9769	-0.9797
1.1634	2.5493	3.6366	4.0000	3.8347	3.3294
4.0000	2.0757	1.0870	0.2301	0.1684	0.0111
0.0406	0.0015	0.0383	1.1706	1.9488	1.9500
-1.8077	0.0000	0.0000	-0.0011	0.0000	-0.9059
2.3942	2.1985	1.9235	1.9718	1.9999	1.9999
4.0000	3.9990	0.0357	0.9286	1.7977	1.8600
0.8623	0.7318	1.2797	1.7869	1.7999	1.8000

Table 5.11. Parameter estimates after each iteration of Q-optimality design for IG1

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
7.5045	13.1446	15.1341	9.4270	16.8146	19.9376
10.2194	20.0000	5.2010	8.7834	9.9160	9.9059
13.5209	11.7701	11.7505	11.4246	10.1925	10.1903
7.3012	16.7123	8.5546	8.0932	8.0421	8.0421
9.8378	9.5516	9.4881	9.5285	9.8706	9.9081
9.0180	10.3185	10.1113	9.8496	9.9349	9.9484
19.4027	5.0859	11.4225	10.8888	9.9068	9.9301
14.0408	10.4813	4.7998	5.0895	11.3919	7.5232
12.8302	19.9675	19.8329	19.9973	18.7889	20.0000
9.2663	19.9986	10.9010	10.8971	9.8442	9.8870
0.0000	0.5592	0.4260	0.6886	0.9352	0.9267
-0.4038	-0.3140	-0.5377	-0.7093	-1.0185	-1.0173
3.5494	0.7252	1.7083	1.8259	1.9935	1.9945
-0.7960	-1.4254	-1.1503	-1.1518	-1.0345	-1.0358
0.0514	0.0002	0.0000	0.0000	0.0000	0.0000
-1.9905	-0.6658	-0.7466	-0.9094	-0.9821	-0.9962
0.0049	3.6583	3.8935	4.0000	4.0000	3.5863
0.0753	1.3525	1.8629	2.1176	2.0561	2.0202
2.7806	0.5754	1.5899	1.8136	1.9230	1.9344
-1.5863	-1.5401	-1.1786	-1.0847	-1.0472	-1.0454
3.4129	1.2302	2.1834	1.9790	1.9910	1.9954
3.9985	0.0049	0.0019	0.0019	0.0005	0.0002
3.9996	4.0000	4.0000	4.0000	4.0000	1.7004

Table 5.12. Parameter estimates after each iteration of MOO design for IG1

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
0.6937	6.3767	4.3934	4.8815	4.9522	4.9579
11.1041	1.0244	3.3500	9.5513	9.9764	9.9762
20.0000	3.1013	17.3311	17.3466	17.3480	17.5984
16.2583	20.0000	8.1641	7.8536	7.8970	7.9124
6.0284	9.2983	10.9735	9.8505	9.8711	9.9060
11.7926	0.0029	9.5921	10.3200	9.9630	9.9643
0.8967	0.0007	9.2258	9.8122	9.9428	9.9047
7.4913	1.5447	11.1497	9.5977	10.0186	10.0127
3.0781	0.0398	0.0000	0.0000	0.0000	0.0000
9.9916	10.5099	20.0000	0.4280	0.4253	0.4244
0.0019	4.0000	2.4434	0.8213	0.9659	0.9781
-1.1819	-1.1746	-0.8050	-0.9550	-0.9952	-0.9985
2.1059	4.0000	1.4341	1.9360	1.9677	1.9965
-0.1017	-1.2450	-0.7575	-0.8536	-0.9864	-0.9852
0.0901	2.4276	1.6703	2.1058	1.9548	1.9595
-1.9127	-0.7059	-0.7851	-0.9595	-0.9811	-0.9981
1.5024	4.0000	1.3399	1.9300	1.9503	1.9505
1.7211	0.6125	0.0396	0.0385	0.0376	0.0295
3.4485	1.3339	1.6810	1.9337	1.9264	1.9500
0.0000	-2.0000	-0.4630	-0.8371	-0.8370	-1.0172
3.1100	1.6786	1.7894	1.9043	1.9924	1.9967
3.9300	3.7257	0.1649	1.7506	1.9846	1.9833
3.8432	3.9338	2.1961	0.5935	0.0355	0.4810

5.4 CONCLUSIONS

This chapter reports an iterative model identification method that tightly integrates identifiability analysis, parameter estimation and model-based design of experiments. The major differences between the proposed iterative procedure and the other existing ones are the two-phase parameter estimation used to estimate AIPs and non-AIPs separately and the methods developed for identifiability analysis and DOE. This iterative procedure is illustrated using a five variable gene regulatory model [175] modeled using S-systems under the BST framework. The results suggested that the

proposed iterative model identification significantly improved the identifiability properties of the model thereby yielding a reliable model. MOO design clearly outperformed the other design criteria. The relative parameter errors were the least for MOO design and this design also yielded parameters with higher number of *a priori* and practical identifiable parameters among the three design criteria considered. This integration has highlighted the fact that, irrespective of the design criteria used, the iterative model identification yields a reliable model in terms of identifiability properties.

CHAPTER 6

CONCLUSIONS & FUTURE WORK

Cells function and survive by orchestrating the expression of genes and their downstream products at the gene, protein and metabolite levels. Of these, metabolites are responsible for much of the functionality of the organism. This motivates the need to comprehend how metabolism works to provide insight into how cells and organism operate. The current advances in technology allow high-throughput experiments at genomic, proteomic and metabolomic levels which can generate time-series data. Typically, pathways are not complicated themselves, but they are highly interconnected since some of the metabolites are coupled with each other through reactions and regulatory interactions. So, it is not easy to predict the behavior and dynamics of metabolism intuitively. This necessitates the use of mathematical modeling for assessing the functioning and regulation of metabolic networks. Usually, these models consist of unknown parameters which need to be estimated by calibrating with experimental data. Although this estimation task has been routinely applied using nonlinear regression, there exist a few but important issues in the estimation of biological kinetic modeling which make such a task a bottleneck in the model building procedure. These issues were discussed at length in Chapter 1.

In the following section, the conclusions obtained from the different chapters of this thesis are presented along with challenges faced, methodology used to tackle these challenges, significant findings and limitations of the findings. The subsequent section addresses potential future works that could be done as an extension of this thesis.

6.1 CONCLUSIONS

The main contribution of this thesis is the parameter-identifiability-centric modeling of biochemical networks using the BST models. At each stage of method development the major focus was to improve the parameter identifiability of the model. The results in each chapter suggest the advantages of integrating identifiability analysis into the model identification cycle to obtain a reliable model. This will also avoid the problem of parameter estimation being ill-posed thereby saving a lot of computational resources in solving the ill-posed problem.

As mentioned in Chapter 1, the contributions to the parameter estimation area from published works in the BST modeling differed in the formulation of the objective function, the optimization algorithms for finding the global minima, and/or the numerical methods for evaluating the objective function. Regardless of the objective function and the numerical algorithms used, a common problem faced in the parameter estimation here and in other modeling exercise is the existence of distinct parameter sets that give similar goodness of fit to experimental data. In essence, such problem is essentially caused by the fact that (1) models are only approximation of the true system and (2) data have limited information from which only a subset of parameters can be identified with sufficient accuracy.

The focus of the first part of this thesis was to tackle the latter of the two causes mentioned above and the aim was to investigate the severity of this problem in the case of BST inverse modeling by developing and applying parameter identifiability analyses (Chapter 2). Two criteria of identifiability are defined and applied: *a priori* and practical.

The analysis methods were based on linear(ized) and nonlinear regression analysis, particularly confidence interval/region of parameter estimates. Finally, two BST models, a GMA model of *L. lactis* lactate production [60] and an S-system model of *E. coli* metabolism [132], were used to demonstrate the inherent problem that is plaguing the inverse modeling of BST. Despite the focus on BST here, the methods developed have general applicability to other model formalisms with an appropriate modification of the identifiability criteria. The results showed that even with noise-free data, it is not possible to completely identify kinetic parameters of metabolic models from typical experiments. Furthermore, when contaminated with noise, the number of practically identifiable parameters dropped even more. Such identifiability problem is the root-cause of the difficulty in getting accurate parameters for kinetic models of metabolism and gives motivation for optimal design of experiments that can generate the most-informative data set for parameter estimation. The methods developed here for identifiability analyses are not restricted to the BST models, but can be extendable to any general nonlinear models. However these methods require initial parameter estimates which maybe unavailable in case of an entirely new pathway/system. As mentioned in Chapter 2, this limitation can be overcome by applying the iterative model identification procedure.

The focus of the next work was to extend the methods developed for the identifiability analyses to alternative BST formalisms, namely, decoupled systems and linlog models (Chapter 3). The methodology used in analyzing identifiability of parameters for decoupled estimation was to derive the appropriate noise propagation of the slope estimation. Once the noise structure was determined, the method described in Chapter 2 were used to perform the identifiability analysis. Although the decoupled

model is derived by converting the ODEs into a set of nonlinear algebraic equations and this suggests that these two models being same, the identifiability analysis revealed the differences between these two models. The decoupled model yielded higher identifiable parameters as compared to the corresponding ODE model. This stems from the fact that in decoupled models there is less correlation among states and parameters.

The other alternative BST formalism that is commonly used is linlog models. Linlog models offer a good alternative of power-law models since rates of the metabolites are linearly dependent on the parameters in linlog models. Many researchers have started using linlog models as an alternative to BST model since it simplifies the task of parameter estimation. So, the focus of this work was to apply the parameter identifiability methods to analyze the linlog models. In the linlog models, the rates are undefined at zero concentration and hence all those time points which had zero concentration were avoided in computing identifiable parameters. This resulted in lesser number of data points for parameter estimation of linlog parameters. The identifiability results suggested that the parameter identifiability of the two linlog models considered were poor due to the basic drawback of linlog models: rate being undefined at zero concentration. The identifiability property would be better for a pathway with no zero concentration data.

The focus of the next work pertained to design of experiments, in which a MBDOE is carried out to maximize information content in the time-series data for parameter identification. The main idea was to design experiments to improve parameter identifiability and at the same time to better the parameter precision. Conventionally, the FIM is used as information metric and some scalar metrics of FIM is maximized to

improve the parameter precision. The use of FIM criteria however has an inherent linear approximation which would yield wrong inferences about the parameter precision when applied to a highly nonlinear model. In this thesis, a new MBDOE was proposed as a multi-objective optimization in which simultaneous minimization of model curvature (as a measure of nonlinearity) and maximization of data informativeness were done. The *in silico* comparison of this MOO design criterion with three other design criteria suggested that (1) methods that account for curvature performed better than FIM-based method and (2) the MOO design performed the best. Although any curvature-based designs face the challenge in the computation of second order sensitivities, accounting for model curvatures is critical in case of highly nonlinear models. The proposed criterion is also flexible in accommodating alternative formulations of curvatures and data informativeness.

Finally, the identifiability analysis and design of experiment method developed were integrated into an iterative procedure for model identification. The particular focus was on the way these individual steps were integrated to improve the parameter precision iteratively. Two-phase parameter estimation was adopted here to separate out the AIPs and non-AIPs. This segregation is indispensible to overcome the ODE stiffness issue arose due to estimating all the parameters including the non-AIPs. This iterative procedure was tested on a five-variable gene regulator network modeled as S-systems. The results supported the fact that iterative model identification incorporating parameter identifiability yields a reliable model thereby avoiding any premature conclusion and wrong inference about the system/process in hand. Another finding was that the MOO design criterion outperformed the other two designs in terms of the number of identifiable

parameters and relative parameter error. Although this iterative model development process was applied to a BST model, this process can very well be extended any other nonlinear dynamic model.

6.2 FUTURE WORK

Chou and Voit [1] have clearly brought out the challenges and problems in the area of parameter estimation and structure identification of genomic and biochemical systems. They classified the problems of inverse modeling into four major issues: data related, model related, computational and mathematical issues. The model selection criteria, the model related issue, was not a problem in this thesis as BST models were of particular focus. Parameter identifiability, which is one of the major causes of computational and mathematical issues, was tackled in this thesis. The following section highlights some of the interesting future topics that could be studied as a continuation of the present thesis.

6.2.1 GLOBAL IDENTIFIABILITY

This thesis has addressed the significance of parameter identifiability to a great extent, but it is worth mentioning that the methods developed for identifiability analyses were local in the sense that it depends on the nominal parameter values However, as mentioned in Chapter 2, there is no reliable global parameter identifiability method for nonlinear model. Currently, there are few software such as DAISY [97], PLE [93] and AMIGO [96] which claim to perform global identifiability analysis. However, these softwares are limited in the size and/or functional form of the non-linearities that can be

handled. Recently, Banga and co-workers have developed a new tool for global identifiability called GenSSI [111], which is implemented as a free toolbox for the MATLAB computing language. This tool is not limited by the functional form of the nonlinearity, but has the same drawbacks as the generating series approach discussed in Section 2.1.5.2. So, working towards a reliable global/structural parameter identifiability method for nonlinear model would be an interesting and important research. This would overcome the drawbacks of the current local identifiability methods proposed in this work.

6.2.2 Parameter Identification of BST Models

The current state of the art in model identifiability analysis gives information whether a particular system is not identifiable from a given dataset. This is valuable information, but the next obvious question would be how to proceed further with this information. The typical way that the identifiability analysis is carried out is by obtaining the Jacobian matrix of the model and incomplete identifiability is then related to the full-rank of this matrix. This information is somewhat similar to saying that there is some too many degrees of freedom. Thus, the first interesting work would be on the lines of the two-phase parameter estimation. The idea is to find a list of parameters whose values if known (known, assumed, or estimated with other means), will make the estimation task identifiable. Alternatively, one can also assign weights for parameters based on the ease of their determination in experiments and the selection of such set of parameters can be formulated as an optimization.

6.2.3 IDENTIFIABILITY ANALYSIS OF RANDOMIZED NETWORKS

Another interesting work to pursue is to carry out identifiability analysis, using the methods developed in this work, on a set of random networks modeled by S-systems. The main question to be answered here is: given the number of components (metabolites) and number of connections, what is the typical number of identifiable parameters? Marino and Voit [48] proposed an automated procedure to extract information which takes advantage of the structural features of S-system model. This procedure can be used to generate large number of networks modeled by S-systems and these candidate models can then be subjected to identifiability analysis. The idea is to plot a distribution of parameter identifiability against the number of components and connections in the system, which would give, on an average, the number of identifiable parameters for the given number of components and connections. The information will be helpful for modelers who are using the S-systems. Although the multiplicative structure of S-systems renders this model generation task easy, the number of combinations grows exponentially ($2^{2n(n+1)}$), where n is the number of dependent variables and this could be a major challenge.

6.2.4 IDENTIFIABILITY AND CHOICE OF MODEL EQUATION

As mentioned in Chapter 1, the model identifiability depends on the model structure, experimental design and the variables measured. While in this thesis, the last two of these factors have been addressed, how much the identifiability of a model is affected by the choice of model equations is still an open question. As described in Chapter 1, mathematical modeling often comprises three steps: (1) experimental

measurements (2) assignment of rate laws to each reaction and (3) parameter calibration with respect to measurements. In each of these steps, the modeler is confronted with many alternative approaches. For example when assigning rate equations, one can choose among power-laws, linlog, Michaelis-Menten, simple mass action kinetics, and many others. The identifiability associated with each choice is not immediately clear and the identifiability analysis methods developed in Chapter 2 can be used to investigate the consequences of choosing one formalism over the other. A part of this work has been addressed in Chapter 3, wherein parameter identifiability of decoupled and linlog models were performed, but further comparison should be done with other choices of rate laws mentioned above.

BIBLIOGRAPHY

1. Chou, I.C. and E.O. Voit, *Recent developments in parameter estimation and structure identification of biochemical and genomic systems.* Mathematical Biosciences, 2009. **219**(2): p. 57-83.
2. Voit, E.O. and J. Almeida, *Decoupling dynamical systems for pathway identification from metabolic profiles.* Bioinformatics, 2004. **20**(11): p. 1670-1681.
3. Klipp, E. *Systems biology in practice : concepts, implementation and application*. 2005; xix:[465 p.].
4. Gunawan.R, K.G. Gadkar, and F.J. Doyle III, *Methods to Identify Cellular Architecture and Dynamics from Experimental Data*, in *Systems Modelling in Cellular Biology: From Concepts to Nuts and Bolts*, Z.Szallasi, J.Stelling, and V.Periwal, Editors. 2006, MIT Press: Cambridge MA.
5. van Riel, N.A.W., *Dynamic modelling and analysis of biochemical networks: Mechanism-based models and model-based experiments.* Briefings in Bioinformatics, 2006. **7**(4): p. 364-374.
6. Box, G.E.P. and N.R. Draper, *Empirical model-building and response surfaces*. Wiley series in probability and mathematical statistics . Applied probability and statistics.1987, New York: : Wiley. xiv, 669 p.
7. Kutalik, Z., K.H. Cho, and O. Wolkenhauer, *Optimal sampling time selection for parameter estimation in dynamic pathway modeling.* BioSystems, 2004. **75**(1-3): p. 43-55.
8. Mendes, P. and D.B. Kell, *Non-linear optimization of biochemical pathways: Applications to metabolic engineering and parameter estimation.* Bioinformatics, 1998. **14**(10): p. 869-883.
9. Rodriguez-Fernandez, M., P. Mendes, and J.R. Banga, *A hybrid approach for efficient and robust parameter estimation in biochemical pathways.* BioSystems, 2006. **83**(2-3 SPEC. ISS.): p. 248-265.
10. Voit, E.O. and J. Schwacke, H, *Understanding through modeling*, in *Systems Biology: Principles , Methods and Concepts*, A.K. Konopka, Editor 2007, CRC Press/Taylor & Francis Books. p. 27-82.
11. Schrodinger, E., *What is life? The physical aspect of the living cell*1945, Cambridge [Eng.]; New York: The University press, The Macmillan company. viii, 91 p.
12. Hodgkin, A.L. and A.F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve.* The Journal of physiology, 1952. **117**(4): p. 500-44.
13. Noble, D., *Cardiac action and pacemaker potentials based on the Hodgkin-Huxley equations.* Nature, 1960. **188**: p. 495-7.
14. Mao, F., et al., *Prediction of Biological Pathways through Data Mining and Information Fusion*, in *Computational methods for understanding bacterial and archaeal genomes*, Y. Xu and J.P. Gogarten, Editors. 2008, Imperial College Press: London. p. 473.

15. Wiener, N., *Cybernetics*1948, New York. 194 p.
16. Turing, A.M., *The Chemical Basis of Morphogenesis.* Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 1952. **237**(641): p. 37-72.
17. Von Neumann, J. and O. Morgenstern, *Theory of games and economic behavior*. [3d ed1953, Princeton: , Princeton University Press. 641 p.
18. Shannon, C.E., *A mathematical theory of communication.* Bell System Technical Journal, 1948. **27**: p. 379-423.
19. Wolkenhauer, O., *Systems biology: the reincarnation of systems theory applied in biology?* Briefings in Bioinformatics, 2001. **2**(3): p. 258-70.
20. Rosen, R., *The representation of biological systems from the standpoint of the theory of categories.* Bulletin of Mathematical Biology, 1958. **20**(4): p. 317-341.
21. Rashevsky, N., *Mathematical biophysics; physico-mathematical foundations of biology*. 3d rev. ed1960, New York. 2 v.
22. Ashby, W.R., *An introduction to cybernetics*1956, New York: , J. Wiley. 295 p.
23. Rosen, R., *Recent Developments in the Theory of Control and Regulation of Cellular Processes*, in *International Review of Cytology*, J.F.D. G.H. Bourne and K.W. Jeon, Editors. 1968, Academic Press. p. 25-88.
24. Goodwin, B.C., *Temporal organization in cells : a dynamic theory of cellular control processes*1963, London ; New York: Academic Press. ix, 163 p.
25. Heinmets, F., *Analysis of normal and abnormal cell growth ; model-system formulations and analog computer studies*1966, New York: , Plenum Press. xiii, 288 p.
26. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* Journal of Molecular Biology, 1961. **3**: p. 318-56.
27. Savageau, M.A., *Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions.* Journal of Theoretical Biology, 1969. **25**(3): p. 365-369.
28. Savageau, M.A., *Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation.* Journal of Theoretical Biology, 1969. **25**(3): p. 370-379.
29. Kacser, H. and J.A. Burns, *The control of flux.* Symp Soc Exp Biol, 1973. **27**: p. 65--104.
30. Curto, R., A. Sorribas, and M. Cascante, *Comparative characterization of the fermentation pathway of Saccharomyces cerevisiae using biochemical systems theory and metabolic control analysis: model definition and nomenclature.* Mathematical Biosciences, 1995. **130**(1): p. 25-50.
31. Savageau, M.A., E.O. Voit, and D.H. Irvine, *Biochemical systems theory and metabolic control theory: 1. fundamental similarities and differences.* Mathematical Biosciences, 1987. **86**(2): p. 127-145.
32. Åström, K.J. and P. Eykhoff, *System identification--A survey.* Automatica, 1971. **7**(2): p. 123-162.
33. Eykhoff, P., *Trends and progress in system identification*. IFAC series for graduates, research workers & practising engineers1981, Oxford ; New York: : Pergamon Press. xvi, 402 p.

34. Ljung, L., *Aspects on the system identification problem.* Signal Processing, 1982. **4**(5-6): p. 445-456.
35. Antoniewicz, M.R., et al., *Metabolic flux analysis in a nonstationary system: Fed-batch fermentation of a high yielding strain of E. coli producing 1,3-propanediol.* Metabolic Engineering, 2007. **9**(3): p. 277-292.
36. Kitano, H., *Computational systems biology.* Nature, 2002. **420**(6912): p. 206-210.
37. Kitano, H., *Systems biology: A brief overview.* Science, 2002. **295**(5560): p. 1662-1664.
38. Carson, E.R. and L. Finkelstein, *Systems identification in biology.* Transactions of the Institute of Measurement and Control, 1982. **4**(4): p. 171-176.
39. Jia, G., G.N. Stephanopoulos, and R. Gunawan, *Parameter estimation of kinetic models from metabolic profiles: two-phase dynamic decoupling method.* Bioinformatics, 2011. **27**(14): p. 1964-1970.
40. Stephanopoulos, G., A.A. Aristidou, and J. Nielsen, *Metabolic engineering*1998, San Diego: Academic Press. xxi, 725 p.
41. Palsson, B., *Systems biology : properties of reconstructed networks*2006, New York: Cambridge University Press. xii, 322 p.
42. Lee, S.Y. and E.T. Papoutsakis, *Metabolic engineering*. Bioprocess technology1999, New York: Marcel Dekker. xxii, 423 p.
43. Gombert, A.K. and J. Nielsen, *Mathematical modelling of metabolism.* Current Opinion in Biotechnology, 2000. **11**(2): p. 180-186.
44. Voit, E.O. and A. Ferreira, *Computational analysis of biochemical systems : a practical guide for biochemists and molecular biologists*2000, Cambridge: Cambridge University Press. xii, 531 p.
45. Hatzimanikatis, V. and J.E. Bailey, *Effects of spatiotemporal variations on metabolic control: Approximate analysis using (log)linear kinetic models.* Biotechnology and Bioengineering, 1997. **54**(2): p. 91-104.
46. *Metabolic engineering*1999, New York :: Marcel Dekker.
47. Marin-Sanguino, A., et al., *Metabolic Engineering with power-law and linear-logarithmic systems.* Mathematical Biosciences, 2009. **218**(1): p. 50-58.
48. Marino, S. and E.O. Voit, *An automated procedure for the extraction of metabolic network information from time series data.* Journal of Bioinformatics and Computational Biology, 2006. **4**(3): p. 665-691.
49. Kimura, S., M. Hatakeyama, and A. Konagaya, *Inference of S-system models of genetic networks from noisy time-series data.* Chem-Bio Informatics Journal, 2004. **4**(1): p. 1-14.
50. Atkinson, M.R., et al., *Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in Escherichia coli.* Cell, 2003. **113**(5): p. 597-607.
51. Vera, J., et al., *Power-law models of signal transduction pathways.* Cellular Signalling, 2007. **19**(7): p. 1531-1541.
52. Hatzimanikatis, V. and J.E. Bailey, *MCA has more to say*. Journal of Theoretical Biology, 1996. **182**(3): p. 233-242.
53. Visser, D. and J.J. Heijnen, *Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics.* Metabolic Engineering, 2003. **5**(3): p. 164-176.

54. Wu, L., et al., *A new framework for the estimation of control parameters in metabolic pathways using lin-log kinetics.* European Journal of Biochemistry, 2004. **271**(16): p. 3348-3359.
55. Heijnen, J.J., *Approximative kinetic formats used in metabolic network modeling.* Biotechnol Bioeng, 2005. **91**(5): p. 534--545.
56. Del Rosario, R.C.H., E. Mendoza, and E.O. Voit, *Challenges in lin-log modelling of glycolysis in Lactococcus lactis.* IET Systems Biology, 2008. **2**(3): p. 136-149.
57. Feng-Sheng Wang, C.-L.K.E.O.V., *Kinetic modeling using S-systems and lin-log approaches.* Biochemical Engineerng, 2007. **33**: p. 238-247.
58. Kresnowati, M.T., W.A. van Winden, and J.J. Heijnen, *Determination of elasticities, concentration and flux control coefficients from transient metabolite data using linlog kinetics.* Metab Eng, 2005. **7**(2): p. 142-53.
59. Savageau, M.A., *Development of fractal kinetic theory for enzyme-catalysed reactions and implications for the design of biochemical pathways.* BioSystems, 1998. **47**(1-2): p. 9-36.
60. Voit, E.O., et al., *Regulation of glycolysis in Lactococcus lactis: an unfinished systems biological case study.* Syst Biol (Stevenage), 2006. **153**(4): p. 286--298.
61. Goel, G., I.C. Chou, and E.O. Voit, *Biological systems modeling and analysis: a biomolecular technique of the twenty-first century.* Journal of biomolecular techniques : JBT, 2006. **17**(4): p. 252-269.
62. Cascante, M., *Comparative characterization of the fermentation pathway of Saccharomyces cerevisiae using biochemical systems theory and metabolic control analysis: Steady-state analysis.* Mathematical Biosciences, 1995. **130**(1): p. 51-69.
63. Torres, N.V., *Modeling approach to control of carbohydrate metabolism during citric acid accumulation by aspergillus niger: I. Model definition and stability of the steady state.* Biotechnology and Bioengineering, 1994. **44**(1): p. 104-111.
64. Torres, N.V., E.O. Voit, and C. Gonzalez-Alcon, *Optimization of nonlinear biotechnological processes with linear programming: Application to citric acid production by Aspergillus niger.* Biotechnol Bioeng, 1996. **49**(3): p. 247-58.
65. Curto, R., et al., *Mathematical models of purine metabolism in man.* Mathematical Biosciences, 1998. **151**(1): p. 1-49.
66. Kikuchi, S., et al., *Dynamic modeling of genetic networks using genetic algorithm and S-system.* Bioinformatics, 2003. **19**(5): p. 643--650.
67. Tsai, K.Y. and F.S. Wang, *Evolutionary optimization with data collocation for reverse engineering of biological networks.* Bioinformatics, 2005. **21**(7): p. 1180-1188.
68. Kim, K.Y., D.Y. Cho, and B.T. Zhang, *Multi-stage evolutionary algorithms for efficient identification of gene regulatory networks*, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*2006. p. 45-56.
69. Tucker, W., Z. Kutalik, and V. Moulton, *Estimating parameters for generalized mass action models using constraint propagation.* Math Biosci, 2007. **208**(2): p. 607--620.

70. Polisetty, P.K., E.O. Voit, and E.P. Gatzke, *Identification of metabolic system parameters using global optimization methods.* Theoretical Biology and Medical Modelling, 2006. **3**.
71. Gonzalez, O.R., et al., *Parameter estimation using simulated annealing for S-system models of biochemical networks.* Bioinformatics, 2007. **23**(4): p. 480-486.
72. Kutalik, Z., W. Tucker, and V. Moulton, *S-system parameter estimation for noisy metabolic profiles using Newton-flow analysis.* IET Systems Biology, 2007. **1**(3): p. 174-180.
73. Marin-Sanguino, A., et al., *Optimization of biotechnological systems through geometric programming.* Theoretical Biology and Medical Modelling, 2007. **4**.
74. Liu, P.K. and F.S. Wang, *Inference of biochemical network models in S-system using multiobjective optimization approach.* Bioinformatics, 2008. **24**(8): p. 1085-1092.
75. Almeida, J.S. and E.O. Voit, *Neural-network-based parameter estimation in S-system models of biological networks.* Genome informatics series : proceedings of the . Workshop on Genome Informatics. Workshop on Genome Informatics, 2003. **14**: p. 114-123.
76. Chou, I.C., H. Martens, and E.O. Voit, *Parameter estimation in biochemical systems models with alternating regression.* Theor Biol Med Model, 2006. **3**: p. 25.
77. Vilela, M., et al., *Parameter optimization in S-system models.* BMC Syst Biol, 2008. **2**: p. 35.
78. Ljung, L., *System identification: theory for the user*. 2nd ed1999, Upper Saddle River, N.J.: Prentice Hall. 609 p.
79. Anderson, J. and A. Papachristodoulou, *On validation and invalidation of biological models.* BMC Bioinformatics, 2009. **10**: p. 132.
80. Beck, M., et al., *On the problem of model validation for predictive exposure assessments.* Stochastic Hydrology and Hydraulics, 1997. **11**(3): p. 229-254.
81. Smith, R.S. and J.C. Doyle, *Model validation: a connection between robust control and identification.* Automatic Control, IEEE Transactions on, 1992. **37**(7): p. 942-952.
82. Feng, X.J. and H. Rabitz, *Optimal Identification of Biochemical Reaction Networks.* Biophysical Journal, 2004. **86**(3): p. 1270-1281.
83. Kremling, A., et al., *A benchmark for methods in reverse engineering and model discrimination: Problem formulation and solutions.* Genome Research, 2004. **14**(9): p. 1773-1785.
84. Yao, K.Z., et al., *Modeling ethylene/butene copolymerization with multi-site catalysts: Parameter estimability and experimental design.* Polymer Reaction Engineering, 2003. **11**(3): p. 563-588.
85. Audoly, S., et al., *Global identifiability of nonlinear models of biological systems.* IEEE Transactions on Biomedical Engineering, 2001. **48**(1): p. 55-65.
86. Vajda, S., K.R. Godfrey, and H. Rabitz, *Similarity transformation approach to identifiability analysis of nonlinear compartmental models.* Mathematical Biosciences, 1989. **93**(2): p. 217-248.
87. Seber, G.A.F. and C.J. Wild, *Nonlinear regression*. Wiley series in probability and statistics.2003, Hoboken, N.J.: Wiley-Interscience. xx, 768 p.

88. Balsa-Canto, E., A.A. Alonso, and J.R. Banga, *An iterative identification procedure for dynamic modeling of biochemical networks.* BMC Syst Biol, 2010. **4**(1): p. 11.
89. Davidescu, F.P. and S.B. JÃ¸rgensen, *Structural parameter identifiability analysis for dynamic reaction networks.* Chemical Engineering Science, 2008. **63**(19): p. 4754-4762.
90. Hengl, S., et al., *Data-based identifiability analysis of non-linear dynamical models.* Bioinformatics, 2007. **23**(19): p. 2612-2618.
91. Jiménez-Hornero, J.E., I.M. Santos-Dueñas, and I. García-García, *Optimization of biotechnological processes. The acetic acid fermentation. Part II: Practical identifiability analysis and parameter estimation.* Biochemical Engineering Journal, 2009.
92. Quaiser, T. and M. Monnigmann, *Systematic identifiability testing for unambiguous mechanistic modeling - application to JAK-STAT, MAP kinase, and NF-kappaB signaling pathway models.* BMC Systems Biology, 2009. **3**(1): p. 50.
93. Raue, A., et al., *Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood.* Bioinformatics, 2009. **25**(15): p. 1923-1929.
94. Roper, R.T., M. Pia Saccomani, and P. Vicini, *Cellular signaling identifiability analysis: A case study.* Journal of Theoretical Biology, 2010. **264**(2): p. 528-537.
95. Yue, H., et al., *Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis: A case study of an NF-ÎºB signalling pathway.* Molecular BioSystems, 2006. **2**(12): p. 640-649.
96. Balsa-Canto, E. and J.R. Banga, *AMIGO, a toolbox for advanced model identification in systems biology using global optimization.* Bioinformatics, 2011. **27**(16): p. 2311-2313.
97. Bellu, G., et al., *DAISY: A new software tool to test global identifiability of biological and physiological systems.* Computer Methods and Programs in Biomedicine, 2007. **88**(1): p. 52-61.
98. Carson, E.R., C. Cobelli, and L. Finkelstein, *The mathematical modeling of metabolic and endocrine systems : model formulation, identification, and validation*. Biomedical engineering and health systems1983, New York: : J. Wiley. xix, 394 p.
99. Bellman, R. and K.J. Åström, *On structural identifiability.* Mathematical Biosciences, 1970. **7**(3-4): p. 329-339.
100. Godfrey, K.R., *The identifiability of parameters of models used in biomedicine.* Mathematical Modelling, 1986. **7**(9-12): p. 1195-1214.
101. JimeÌ • nez-Hornero, J.E., I.M. Santos-DueniÌ*f*as, and I. GarciÌ • a-GarciÌ • a, *Structural identifiability of a model for the acetic acid fermentation process.* Mathematical Biosciences, 2008. **216**(2): p. 154-162.
102. Nikerel, I.E., et al., *Model reduction and a priori kinetic parameter identifiability analysis using metabolome time series for metabolic reaction networks with linlog kinetics.* Metab Eng, 2008.
103. Faller, D., U. Klingmüller, and J. Timmer, *Simulation Methods for Optimal Experimental Design in Systems Biology.* Simulation, 2003. **79**(12): p. 717-725.

104. Chappell, M.J. and K.R. Godfrey, *Structural identifiability of the parameters of a nonlinear batch reactor model.* Mathematical Biosciences, 1992. **108**(2): p. 241-251.
105. Chappell, M.J., K.R. Godfrey, and S. Vajda, *Global identifiability of the parameters of nonlinear systems with specified inputs: A comparison of methods.* Mathematical Biosciences, 1990. **102**(1): p. 41-73.
106. K.R.Godfrey and I. J.J.DiStefano, *Identifiability of model parameters*, in *Identifiability of Parametric Models*, E.Walter, Editor 1987, Pergamon: Oxford.
107. Y.Lecourtier, F.Lamnabhi-Lagarrigue, and E.Walter, *Volterra and Generating Power Series approaches to Identifiability testing*, in *Identifiability of Parametric Models*, E.Walter, Editor 1987, Pergamon: Oxford.
108. Walter, E. and L. Pronzato, *On the identifiability and distinguishability of nonlinear parametric models.* Mathematics and Computers in Simulation, 1996. **42**(2-3): p. 125-134.
109. Davidescu, F.P. and S.B. Jørgensen, *Structural parameter identifiability analysis for dynamic reaction networks.* Chemical Engineering Science, 2008. **63**(19): p. 4754-4762.
110. Savageau, M.A., *Design principles for elementary gene circuits: Elements, methods, and examples.* Chaos, 2001. **11**(1): p. 142-159.
111. Chis, O., J.R. Banga, and E. Balsa-Canto, *GenSSI: a software toolbox for structural identifiability analysis of biological models.* Bioinformatics, 2011.
112. Ljung, L. and T. Glad, *On global identifiability for arbitrary model parametrizations.* Automatica, 1994. **30**(2): p. 265-276.
113. Denis-Vidal, L., G. Joly-Blanchard, and C. Noiret, *Some effective approaches to check the identifiability of uncontrolled nonlinear systems.* Mathematics and Computers in Simulation, 2001. **57**(1-2): p. 35-44.
114. Müller, T.G., *Modeling complex systems with differential equations*, 2002, Albert-Ludwigs Universität Freiburg: Breisgau.
115. Margaria, G., et al., *Differential algebra methods for the study of the structural identifiability of rational function state-space models in the biosciences.* Mathematical Biosciences, 2001. **174**(1): p. 1-26.
116. Chappell, M.J. and R.N. Gunn, *A procedure for generating locally identifiable reparameterisations of unidentifiable non-linear systems by the similarity transformation approach.* Mathematical Biosciences, 1998. **148**(1): p. 21-41.
117. Braems, I., et al. *Guaranteed numerical alternatives to structural identifiability testing*. in *Proceedings of the IEEE Conference on Decision and Control*. 2001.
118. Walter, E., et al., *Guaranteed Numerical Computation as an Alternative to Computer Algebra for Testing Models for Identifiability*, in *Numerical Software with Result Verification*2004. p. 563-577.
119. Sedoglavic, A., *A probabilistic algorithm to test local algebraic observability in polynomial time.* Journal of Symbolic Computation, 2002. **33**(5): p. 735-755.
120. Ingalls, B., *Sensitivity analysis: from model parameters to system behaviour.* Essays in biochemistry, 2008. **45**: p. 177-193.
121. Turaì • nyi, T., *Sensitivity analysis of complex kinetic systems. Tools and applications.* Journal of Mathematical Chemistry, 1990. **5**(3): p. 203-248.

122. Varma, A., M. Morbidelli, and H. Wu, *Parametric sensitivity in chemical systems*. Cambridge series in chemical engineering.1999, Cambridge, U.K. ; New York, NY: Cambridge University Press. xvi, 342 p.
123. Zak, D.E., et al., *Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: Insights from an identifiability analysis of an in silico network.* Genome Research, 2003. **13**: p. 2396-2405.
124. Voit, E.O., S. Marino, and R. Lall, *Challenges for the identification of biological systems from in vivo time series data.* In Silico Biology, 2005. **5**(2): p. 83-92.
125. Seber, G.A.F. and C.J. Wild, *Nonlinear regression*. Wiley series in probability and mathematical statistics . Probability and mathematical statistics.1989, New York: : Wiley. xx, 768 p.
126. Vilela, M., et al., *Identification of neutral biochemical network models from time series data.* BMC Systems Biology, 2009. **3**(1): p. 47.
127. Beck, J.V. and K.J. Arnold, *Parameter estimation in engineering and science*1977, New York: : Wiley. xix, 501 p.
128. Emery, A.F. and A.V. Nenarokomov, *Optimal experiment design.* Measurement Science and Technology, 1998. **9**(6): p. 864-876.
129. Landaw, E.M. and J.J. DiStefano 3rd, *Multiexponential, multicompartmental, and noncompartmental modeling. II. Data analysis and statistical considerations.* The American journal of physiology, 1984. **246**(5 Pt 2): p. R665-677.
130. Boyd, S., et al., *Linear Matrix Inequality in System and Control Theory*1994, Philadephia: Society for Industrial and Applied Mathematics. 193.
131. Gadkar, K.G., R. Gunawan, and F.J. Doyle, *Iterative approach to model identification of biological networks.* BMC Bioinformatics, 2005. **6**: p. 155.
132. Ko, C.L., et al., *S-system approach to modeling recombinant Escherichia coli growth by hybrid differential evolution with data collocation.* Biochemical Engineering Journal, 2006. **28**(1): p. 10-16.
133. Bolotin, A., et al., *The complete genome sequence of the lactic acid bacterium lactococcus lactis ssp. lactis IL1403.* Genome Research, 2001. **11**(5): p. 731-753.
134. De Vos, W.M., *Safe and sustainable systems for food-grade fermentations by genetically modified lactic acid bacteria.* International Dairy Journal, 1999. **9**(1): p. 3-10.
135. Neves, A.R., et al., *Effect of different NADH oxidase levels on glucose metabolism by Lactococcus lactis: kinetics of intracellular metabolite pools determined by in vivo nuclear magnetic resonance.* Appl Environ Microbiol, 2002. **68**(12): p. 6332--6342.
136. Neves, A.R., et al., *In vivo nuclear magnetic resonance studies of glycolytic kinetics in Lactococcus lactis.* Biotechnol Bioeng, 1999. **64**(2): p. 200--212.
137. Balsa-Canto, E., A.A. Alonso, and J.R. Banga, *Computational procedures for optimal experimental design in biological systems.* IET Systems Biology, 2008. **2**(4): p. 163-172.
138. Seatzu, C., *A fitting based method for paramter estimation in S-systems.* Dynamic System Application, 2000: p. 77.
139. Nau, B., *Averaging and exponential smoothing models*, in *University of Duke, Course webpage*2005.

140. Draper, N.R. and H. Smith, *Applied Regression Analysis*1998, New York: Wiley-Interscience.
141. Wang, F.-S., C.-L. Ko, and E.O. Voit, *Kinetic modeling using S-systems and lin-log approaches.* Biochemical Engineering Journal, 2007. **33**(3): p. 238-247.
142. Chou, I.C. and E.O. Voit, *Parameter Estimation in Canonical Biological System Models.* International Journal of Systems and Synthetic Biology, 2010. **1**(1): p. 1-19.
143. Franceschini, G. and S. Macchietto, *Model-based design of experiments for parameter precision: State of the art.* Chemical Engineering Science, 2008. **63**(19): p. 4846-4872.
144. Smith, K., *On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and its Constants and the Guidance they give Towards a Proper Choice of the Distribution of Observations.* Biometrika, 1918. **12**(1/2): p. 1-85.
145. Kiefer, J. and J. Wolfowitz, *Optimum Designs in Regression Problems.* The Annals of Mathematical Statistics, 1959. **30**(2): p. 271-294.
146. Box, G.E.P. and H.L. Lucas, *Design of Experiments in Non-Linear Situations.* Biometrika, 1959. **46**(1/2): p. 77-90.
147. Cramer, H., *Mathematical methods of statistics*1946, Princeton: , Princeton university press. xvi, 575 p. incl. tables, diagrs.
148. Atkinson, A.C. and B. Bogacka, *Compound and other optimum designs for systems of nonlinear differential equations arising in chemical kinetics.* Chemometrics and Intelligent Laboratory Systems, 2002. **61**(1-2): p. 17-33.
149. Wald, A., *On the Efficient Design of Statistical Investigations.* The Annals of Mathematical Statistics, 1943. **14**(2): p. 134-140.
150. Chernoff, H., *Locally Optimal Designs for Estimating Parameters.* The Annals of Mathematical Statistics, 1953. **24**(4): p. 586-602.
151. Ehrenfeld, S., *On the Efficiency of Experimental Designs.* The Annals of Mathematical Statistics, 1955. **26**(2): p. 247-255.
152. Shirt, R.W., T.J. Harris, and D.W. Bacon, *Experimental Design Considerations for Dynamic Systems.* Industrial & Engineering Chemistry Research, 1994. **33**(11): p. 2656-2667.
153. Goodwin, G.C. and R.L. Payne, *Dynamic system identification : experiment design and data analysis*. Mathematics in science and engineering1977, New York: : Academic Press. x, 291 p.
154. Franceschini, G. and S. Macchietto, *Validation of a Model for Biodiesel Production through Model-Based Experiment Design.* Industrial & Engineering Chemistry Research, 2006. **46**(1): p. 220-232.
155. Franceschini, G. and S. Macchietto, *Novel anticorrelation criteria for model-based experiment design: Theory and formulations.* AIChE Journal, 2008. **54**(4): p. 1009-1024.
156. Van Derlinden, E., K. Bernaerts, and J.F. Van Impe, *Simultaneous versus sequential optimal experiment design for the identification of multi-parameter microbial growth kinetics as a function of temperature.* Journal of Theoretical Biology, 2010. **264**(2): p. 347-355.

157. Walter, E. and L. Pronzato, *Qualitative and quantitative experiment design for phenomenological models - A survey.* Automatica, 1990. **26**(2): p. 195-213.
158. Benabbas, L., S.P. Asprey, and S. Macchietto, *Curvature-Based Methods for Designing Optimally Informative Experiments in Multiresponse Nonlinear Dynamic Situations.* Industrial & Engineering Chemistry Research, 2005. **44**(18): p. 7120-7131.
159. Bogacka, B. and F. Wright, *Comparison of Two Design Optimality Criteria Applied to a Nonlinear Model.* Journal of Biopharmaceutical Statistics, 2004. **14**(4): p. 909-930.
160. Merlé, Y. and M. Tod, *Impact of pharmacokinetic-pharmacodynamic model linearization on the accuracy of population information matrix and optimal design.* J Pharmacokinet Pharmacodyn, 2001. **28**(4): p. 363-88.
161. Cochran, W.G., *Experiments for Nonlinear Functions.* Journal of the American Statistical Association, 1973. **68**(344): p. 771-781.
162. Box, M.J., *Bias in Nonlinear Estimation.* Journal of the Royal Statistical Society. Series B (Methodological), 1971. **33**(2): p. 171-201.
163. Clarke, G.P.Y., *Moments of the Least Squares Estimators in a Non-Linear Regression Model.* Journal of the Royal Statistical Society. Series B (Methodological), 1980. **42**(2): p. 227-237.
164. Bates, D.M. and D.G. Watts, *Relative Curvature Measures of Nonlinearity.* Journal of the Royal Statistical Society. Series B (Methodological), 1980. **42**(1): p. 1-25.
165. Hamilton, D.C. and D.G. Watts, *A Quadratic Design Criterion for Precise Estimation in Nonlinear Regression Models.* Technometrics, 1985. **27**(3): p. 241-250.
166. O'Brien, T.E., *A Note on Quadratic Designs for Nonlinear Regression Models.* Biometrika, 1992. **79**(4): p. 847-849.
167. Guay, M., *Curvature measures for multiresponse regression models.* Biometrika, 1995. **82**(2): p. 411-417.
168. Bates, D.M. and D.G. Watts, *Nonlinear regression analysis and its applications*. Wiley series in probability and mathematical statistics . Applied probability and statistics.1988, New York: : Wiley. xiv, 365 p.
169. Rangaiah, G.P. and World Scientific (Firm). *Multi-objective optimization : techniques and applications in chemical engineering*. Advances in process systems engineering v. 1. 2009; xvii, 435 p.].
170. Vafaeyan, S. and J. Thibault, *Selection of pareto-optimal solutions for process optimization using rough set method: A new approach.* Computers & Chemical Engineering, 2009. **33**(11): p. 1814-1825.
171. Franceschini, G. and S. Macchietto, *Anti-Correlation Approach to Model-Based Experiment Design: Application to a Biodiesel Production Process.* Industrial & Engineering Chemistry Research, 2008. **47**(7): p. 2331-2348.
172. Noureddini, H. and D. Zhu, *Kinetics of transesterification of soybean oil.* Journal of the American Oil Chemists' Society, 1997. **74**(11): p. 1457-1463.
173. Asprey, S.P. and S. Macchietto, *Statistical tools for optimal dynamic model building.* Computers & Chemical Engineering, 2000. **24**(2-7): p. 1261-1267.

174. Maki, Y., et al., *Inference of genetic network using the expression profile time course data of mouse P19 cells.* Genome Informatics, 2002. **13**: p. 382-383.
175. Hlavacek, W.S. and M.A. Savageau, *Rules for coupled expression of regulator and effector genes in inducible circuits.* Journal of Molecular Biology, 1996. **255**(1): p. 121-139.
176. Voit, E.O., *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*, E.O. Voit, Editor 2000, Cambridge University Press. p. 222-259.

APPENDIX A

Non-linear Regression

Consider a non-linear model given by,

$$\mathbf{y} = f(\mathbf{X};\boldsymbol{\theta}) + \boldsymbol{\varepsilon}. \tag{A.1}$$

The generalized least squares estimator of the true minimum θ^* is $\hat{\theta}_G$, which minimizes the error sum of squares:

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y_i - f(x_i;\boldsymbol{\theta})\right)^T W \left(y_i - f(x_i;\boldsymbol{\theta})\right) \tag{A.1}$$

Depending upon the weight *W*, the least squares estimation can be classified into three cases.

General Least Squares

Case A	**Case B**	**Case C**
W=I	W=V^{-1}	W=V^{-1}
A1: V= σ^2I	**B1:** V= σ^2I	**C1:** V= σ^2I
A2: V= $\sigma^2\Omega$	**B2**: V= $\sigma^2\Omega$	**C2**: V= $\sigma^2\Omega$
A3: V	**B3**: V	**C3**: V

Case A: W=I, Ordinary Least Squares (OLS)

$$\text{cov}\left(\hat{\boldsymbol{\theta}}\right) = \left(\mathbf{S^T S}\right)^{-1} \mathbf{S^T V S} \left(\mathbf{S^T S}\right)^{-1} \quad \textbf{since } \hat{\boldsymbol{\beta}} = \left(\mathbf{S^T S}\right)\mathbf{S^T}\varepsilon$$

$$\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T \left(\mathbf{S^T S}\right)\left(\mathbf{S^T V S}\right)^{-1}\left(\mathbf{S^T S}\right)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) < \chi^2_{p,\alpha}$$

A1: V= σ^2I

$$\text{cov}\left(\hat{\boldsymbol{\theta}}\right) = \left(\mathbf{S^T S}\right)^{-1}\sigma^2 \Rightarrow \sigma^{-2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T \left(\mathbf{S^T S}\right)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) < \chi^2_{p,\alpha}$$

A2: V= $\sigma^2\Omega$, σ^2 unknown and Ω known

$$\text{cov}\left(\hat{\boldsymbol{\theta}}\right)=\sigma^2\left(\mathbf{S^TS}\right)^{-1}\mathbf{S^T\Omega S}\left(\mathbf{S^TS}\right)^{-1}$$

$$\Rightarrow\sigma^{-2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{S^TS}\right)\left(\mathbf{S^T\Omega S}\right)^{-1}\left(\mathbf{S^TS}\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)<\chi^2_{p,\alpha}$$

Case B: W=V^{-1}

$$\text{cov}\left(\hat{\boldsymbol{\theta}}\right)=\left(\mathbf{S^TV^{-1}S}\right)^{-1}\Rightarrow\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{S^TV^{-1}S}\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)<\chi^2_{p,\alpha}$$

B1: V= σ^2I

$$\text{cov}\left(\hat{\boldsymbol{\theta}}\right)=\sigma^2\left(\mathbf{S^TV^{-1}S}\right)^{-1}\Rightarrow\sigma^{-2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{S^TV^{-1}S}\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)<\chi^2_{p,\alpha}$$

B2: V= $\sigma^2\Omega$, σ^2 unknown and Ω known

$$\text{cov}\left(\hat{\boldsymbol{\theta}}\right)=\sigma^2\left(\mathbf{S^TV^{-1}S}\right)^{-1}\Rightarrow\sigma^{-2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{S^TV^{-1}S}\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)<\chi^2_{p,\alpha}$$

Case C: W=W

$$\text{cov}\left(\hat{\boldsymbol{\theta}}\right)=\sigma^2\left(\mathbf{S^TWS}\right)^{-1}\left(\mathbf{S^TWVWS}\right)\left(\mathbf{S^TWS}\right)^{-1}$$

$$\Rightarrow\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{S^TWS}\right)\left(\mathbf{S^TWVWS}\right)^{-1}\left(\mathbf{S^TWS}\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)<\chi^2_{p,\alpha}$$

C1: V= σ^2I

$$\text{cov}\left(\hat{\boldsymbol{\theta}}\right)=\sigma^2\left(\mathbf{S^TWS}\right)^{-1}\left(\mathbf{S^TWWS}\right)\left(\mathbf{S^TWS}\right)^{-1}$$

$$\Rightarrow\sigma^{-2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{S^TWS}\right)\left(\mathbf{S^TWWS}\right)^{-1}\left(\mathbf{S^TWS}\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)<\chi^2_{p,\alpha}$$

C2: V= $\sigma^2\Omega$

$$\text{cov}\left(\hat{\boldsymbol{\theta}}\right)=\sigma^2\left(\mathbf{S^TWS}\right)^{-1}\left(\mathbf{S^TW\Omega WS}\right)\left(\mathbf{S^TWS}\right)^{-1}$$

$$\Rightarrow\sigma^{-2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{S^TWS}\right)\left(\mathbf{S^TW\Omega WS}\right)^{-1}\left(\mathbf{S^TWS}\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)<\chi^2_{p,\alpha}$$

Legend:

W - Weighting matrix

V - covariance matrix

σ - Standard deviation

Ω - Structure of covariance matrix

S - Sensitivity matrix

Additional Information for Case Studies

a. Case Study 1:

Table A below gives the detailed identifiability property of the parameters in the *L. lactis* model. Method 2 and 3 are only applied to *a priori* identifiable parameters (AIP).

Table A: Parametric Identifiability in *L. lactis* Model

Parameter	Parameter value	AIP*	Practical Identifiability		
			Method 1	Method 2	Method 3
α_1	0.3592	x	–		
β_1	0.3115	✓	–	x	x
α_3	1.1452	x	–		
β_{51}	0.0417	x	–		
β_2	0.4698	✓	–	x	x
β_3	2.167	x	–		
β_{41}	0.9375	✓	–	x	x
β_{42}	0.2087	x	–		
β_{52}	1.3258	✓	–	x	x
g_{1Glc}	1.1287	✓	–	✓	✓
g_{11}	-1.2906	x	–		
g_{25}	0.2168	✓	–	x	x

h_{11}	2.17	✓	–	✓	✓
g_{34}	3.5453	✓	–	x	x
h_{1atp}	0.8152	✓	–	x	x
h_{22}	1.0297	✓	–	✓	✓
h_{2P}	0.2377	✓	–	x	x
h_{33}	2.1649	✓	–	x	x
h_{412}	0.8744	✓	–	x	x
h_{414}	0.0991	x	–		
h_{41P}	-0.0005	x	–		
h_{424}	0.0002	x	–		
h_{515}	0.6202	x	–		
h_{512}	0.9263	✓	–	✓	✓
h_{525}	1.5255	✓	–	✓	✓

* ✓ indicates that the parameter is **identifiable** and x indicates the parameter is **not identifiable**

b. **Case Study 2:**

Table B gives the detailed identifiability property of the parameters in the above model. Method 2 and 3 are only applied to AIP.

Table B: Parametric Identifiability in *E. coli* Model

Parameter	Parameter value	AIP	Practical Identifiability		
			Method 1	Method 2	Method 3
α_1	0.4973	✓	✓	✓	✓
α_2	0.0817	x	x		
α_3	0.2858	✓	✓	x	✓
α_4	3.7124	✓	✓	✓	✓
α_5	0.4562	✓	✓	x	✓
β_2	1.2484	✓	✓	x	✓
β_3	0.1285	x	x		

β_4	2.5318	x	x		
β_5	0.0335	x	x		
g_{11}	0.9099	✓	x	✓	✓
g_{12}	0.1301	✓	x	x	x
g_{31}	0.7366	✓	x	✓	✓
g_{32}	0.1311	x	x		
g_{41}	1.7076	✓	x	x	x
g_{42}	0.1252	✓	x	x	x
g_{51}	0.2292	x	x		
g_{52}	0.0277	x	x		
h_{11}	1.7514	✓	x	✓	✓
h_{12}	0.1292	x	x		
h_{21}	0.9325	x	x		
h_{22}	0.1927	✓	✓	✓	✓
h_{31}	1.3535	✓	x	✓	✓
h_{32}	0.1175	x	x		
h_{33}	-0.011	x	x		
h_{41}	1.9875	✓	x	x	x
h_{42}	0.121	x	x		
h_{44}	-0.01	x	x		
h_{51}	1.1975	✓	x	x	✓
h_{52}	0.4462	x	x		
h_{55}	-0.0426	x	x		

* ✓ indicates that the parameter is **identifiable** and x indicates the parameter is **not identifiable**

APPENDIX B

Iterative model identification: This section presents the extension of the results presented in Section 5.3.1. The results in aforementioned section pertain only to IG1. The following set of tables present the summary of identifiability results, AIPs and relative parameter errors after each iteration for IG2 and IG3. The estimates are presented for the three designs: D-optimality, Q-optimality and MOO, for three initial guesses for parameters.

Table B.1. Summary of identifiability results at the end of iterative model identification for IG2

Design	AIP	PIP
D-optimality	14/23	11/14
Q-optimality	16/23	12/16
MOO	17/23	14/17

Table B.2. AIP after each iteration for D-optimality design and IG2

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
2	3	2	1	1	1
3	13	9	2	2	2
15	15	15	3	8	8
17	17	17	9	9	9
18	18	18	11	11	11
23	23	21	14	12	12
		23	15	13	13
			16	14	14
			17	15	15
			18	16	16
			21	17	17
			23	18	18
				21	21
				23	23

Table B.3. Parameter estimates after each iteration of D-optimality design for IG2

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
16.6167	19.9958	4.3986	4.6549	5.0426	5.0300
17.1753	20.0000	14.0712	11.3375	10.0043	10.0350
14.8432	15.8277	14.0601	14.1856	11.8525	10.7793
20.0000	16.9499	15.5394	20.0000	15.3699	20.0000
8.0944	5.8742	5.2723	0.0029	0.0683	0.0116
2.6699	9.1256	11.2195	12.9662	13.0667	18.0398
19.9927	20.0000	20.0000	20.0000	19.0101	19.9739
18.9328	19.9995	20.0000	13.7422	10.9013	10.8971
7.7140	19.9971	19.3370	10.7021	10.3923	10.2966
19.9946	19.9990	20.0000	20.0000	20.0000	20.0000
0.0020	3.7590	2.5183	0.6749	0.8550	0.9036
-1.9986	-1.9999	-1.9918	-1.9997	-0.8632	-0.8937
3.8549	0.1482	2.2298	2.1077	2.1330	2.1327
-1.6350	-1.9963	-1.2360	-1.0867	-1.0304	-1.0365
0.2216	0.6854	1.6755	1.8501	1.9096	1.9143
-2.0000	-1.8044	-1.2442	-1.1691	-0.9716	-0.9716
3.1512	2.7005	2.2046	2.1339	2.0892	2.0803
3.9270	3.9939	2.9696	2.1746	2.1176	2.1179
2.2533	0.0020	0.0020	0.0000	0.0040	0.0818
-0.0013	0.0000	-0.3992	-0.2841	-0.2897	-0.2106
3.4850	3.2415	2.2405	2.2405	2.1239	2.1299
2.7015	0.0023	0.0023	0.0001	0.0000	0.0001
2.7833	1.7995	1.8347	1.9753	1.9753	1.9761

Table B.4. AIP after each iteration for Q-optimality design and IG2

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
6	2	2	2	2	2
7	4	5	5	5	6
9	5	6	6	6	8
12	6	9	8	8	9
21	12	11	9	9	10
	21	12	11	11	11
		13	12	12	12
		15	13	13	13
		17	15	15	15
		20	17	16	16
		21	18	17	17
			20	18	18
			21	20	20
			23	21	21
				22	22
				23	23

Table B.5. Parameter estimates after each iteration of Q-optimality design for IG2

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
19.9999	0.1043	0.4927	6.5517	12.4399	4.7962
20.0000	10.7826	10.1729	10.6148	9.9188	9.9490
11.6094	12.7114	11.0354	9.8753	8.6311	7.8989
20.0000	20.0000	0.5534	1.0603	5.3209	11.5770
20.0000	18.4805	7.7652	9.6406	10.0762	16.3767
7.3285	10.4251	9.6117	9.6247	9.8943	9.8832
12.6888	20.0000	20.0000	20.0000	20.0000	20.0000
13.5015	20.0000	10.6484	10.9013	10.9010	10.8971
8.3353	20.0000	2.2097	8.4769	9.9134	9.9484
16.2455	20.0000	14.7429	19.4986	12.7474	10.0934
0.2337	1.5048	0.7197	0.8063	0.8962	0.9494
-0.8966	-0.9031	-0.9218	-0.9799	-0.9843	-0.9862
0.0044	4.0000	1.2429	1.0800	1.7610	1.8361
-1.8809	-1.9992	-1.9930	-1.9989	-1.9991	-1.9996
1.7073	1.7031	1.7265	1.7523	1.9143	1.9158
-1.8591	-1.9956	-0.6658	-0.7696	-0.8956	-0.9348
0.0000	3.8435	1.6462	1.6726	1.9180	1.9159
2.8554	0.2343	2.1176	1.6108	1.8161	1.8629
4.0000	0.0002	0.0017	0.0017	0.0017	0.0007
-0.7306	-1.1154	-0.8341	-0.8407	-0.9012	-0.9088
1.4828	1.5160	1.6137	1.6525	2.1780	2.1797
3.9062	0.0001	0.0001	1.0872	1.9845	1.9820
4.0000	2.4645	1.6408	1.7660	2.0123	2.0477

Table B.6. AIP after each iteration for MOO design and IG2

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
4	4	5	4	2	2
5	5	6	5	4	4
13	13	8	6	5	5
18	18	9	8	6	6
21	19	10	9	8	8
	21	12	10	9	9
		13	12	12	12
		15	13	13	13
		16	15	14	14
		17	16	15	15
		18	17	16	16
		19	18	17	17
		20	19	18	18
		21	20	19	19
			21	20	20
				21	21
				23	23

Table B.7. Parameter estimates after each iteration of MOO design for IG2

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
14.2568	19.9986	19.9995	19.9999	19.9979	20.0000
20.0000	15.3164	14.6462	9.6148	9.7826	9.7857
19.9999	9.0774	6.3670	4.0066	3.9099	3.5115
5.2188	3.2173	6.3330	7.8135	7.8455	7.9988
15.1844	5.0714	2.7503	2.8684	9.8237	9.8237
20.0000	19.6005	13.4690	8.4952	10.3164	10.3853
3.7624	13.6366	0.3076	0.0378	0.0046	0.0957
12.4450	20.0000	8.1206	9.8215	9.3709	9.3854
1.5496	2.0378	10.8762	9.4764	9.7436	9.7272
17.4182	20.0000	1.4287	1.5876	8.6149	12.9885
0.0110	0.0023	0.0023	0.0065	0.0017	0.0007
-0.7464	-0.3530	-0.5776	-0.5760	-1.0811	-1.0487
2.7371	2.2298	2.1077	2.1330	2.1327	2.0537
-0.9937	-1.9996	-1.9996	-1.9988	-1.0702	-1.0565
1.4441	1.4441	1.5371	1.5593	1.7505	1.8198
-1.9986	-1.1976	-0.6927	-1.1976	-0.9198	-0.9691
3.5500	3.9919	2.8574	2.6503	2.0818	2.0818
2.2539	2.1742	2.0990	2.1312	2.0536	2.0481
3.9996	3.6909	3.2808	2.2115	1.9188	1.9356
0.0000	-1.9922	-1.9504	-1.9973	-0.9372	-0.9522
3.5900	0.3835	3.4095	3.0243	2.1423	2.1187
3.9976	0.0230	0.0035	0.0015	0.0049	0.0030
4.0000	0.8323	2.6888	2.5244	2.1393	2.0851

Table B.8. Summary of identifiability results at the end of iterative model identification for IG3

Design	AIP	PIP
D-optimality	14/23	11/15
Q-optimality	14/23	10/14
MOO	16/23	14/16

Table B.9. AIP after each iteration for D-optimality design and IG3

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
2	2	5	2	2	2
8	8	7	5	7	7
14	14	8	8	8	8
15	15	13	13	9	9
17	17	14	14	10	10
18	18	15	15	13	13
22	21	18	18	15	15
	22	21	21	18	17
		22	22	21	18
					21

Table B.10. Parameter estimates after each iteration of D-optimality design for IG3

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
14.3171	19.5893	12.1508	19.3768	12.0080	19.4923
2.7266	6.9477	19.9809	12.3449	9.6547	9.7290
12.1007	7.8858	7.1240	6.0361	7.2632	5.4867
20.0000	20.0000	20.0000	20.0000	19.9995	19.9975
20.0000	19.9461	17.4098	15.7309	10.0280	18.7531
3.7018	11.9292	5.2181	10.6262	0.0061	8.6988
20.0000	18.8171	4.4292	11.8468	10.2863	10.1989
4.0731	8.1676	8.1862	9.1538	9.2592	9.2494
19.9953	19.9976	20.0000	12.6172	10.9188	10.9294
19.7846	16.2455	19.4986	13.2957	10.7474	10.6953
0.0136	3.9960	3.6093	3.6857	3.5674	0.0375
-1.8478	-1.9884	-1.9994	-1.9984	-1.9984	-1.9992
0.9997	4.0000	1.4617	2.2088	2.2281	2.2142
-0.0108	-1.1940	-1.1950	-1.0202	-1.9936	-1.9982
0.0188	0.0310	1.0250	1.9581	1.9581	1.9581
-1.9998	-1.9999	-2.0000	-2.0000	-2.0000	-2.0000
1.2381	3.0730	1.5523	1.7553	1.8315	1.8253
3.8433	1.1485	1.5296	1.8508	1.8439	1.8497
3.9959	0.0100	0.0013	0.0013	0.0027	0.0037
-0.2211	-0.3458	-0.6222	-0.4191	-0.7642	-0.4535
3.2346	2.2388	2.4577	2.2268	2.0502	2.0327
3.8755	3.8151	3.7205	0.0005	0.0002	0.0012
3.7573	4.0000	4.0000	4.0000	4.0000	3.9997

Table B.11. AIP after each iteration for Q-optimality design and IG3

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
5	5	2	2	2	2
13	6	5	5	4	4
15	13	6	6	5	5
18	15	12	7	6	6
21	18	13	12	7	7
	21	15	13	12	12
		16	15	13	13
		18	16	15	15
		21	17	16	16
		22	18	17	17
			21	18	18
			22	21	21
			23	22	22
				23	23

Table B.12. Parameter estimates after each iteration of Q-optimality design for IG3

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
16.8480	19.7324	20.0000	20.0000	20.0000	1.6713
20.0000	19.9913	17.7337	9.5207	9.8007	9.8059
19.9882	20.0000	14.4036	15.6550	15.5612	7.0524
20.0000	20.0000	20.0000	1.7924	8.7580	8.6386
16.6410	16.6889	8.1222	10.9411	10.2451	10.1738
14.5530	16.5152	13.5950	10.7507	9.8266	9.8247
10.6871	16.0727	20.0000	10.5348	10.5717	10.5248
1.2168	4.2593	4.2673	0.2663	0.0989	0.2687
6.0208	7.9040	6.9671	6.9478	1.1065	4.5632
9.2189	13.1766	19.2589	19.8071	19.8071	4.9755
0.0553	1.3558	1.2776	1.2154	1.7582	1.7642
-1.9900	-1.9998	-1.2779	-0.7195	-0.9367	-0.9333
2.2596	1.5709	1.7001	1.8639	1.8792	1.8799
-1.9581	-1.9999	-1.9886	-1.9984	-1.9874	-2.0000
0.3472	1.2793	1.8143	2.1329	1.8001	1.8600
-2.0000	-2.0000	-1.2681	-0.6359	-1.2779	-0.9805
0.0003	3.3622	1.3491	1.4358	1.8470	1.9150
2.2539	3.0691	2.0990	2.1312	2.1536	2.1481
3.9999	0.0022	0.0000	0.0000	0.0000	0.6602
-0.0437	-0.3860	-0.8770	-0.5610	-0.5610	-1.5124
2.4345	2.8348	2.8952	2.8315	2.2402	2.1500
4.0000	4.0000	4.0000	2.6409	2.1877	2.1771
4.0000	0.2410	1.3286	2.1395	1.8826	1.9455

Table B.13. AIP after each iteration for MOO design and IG3

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
2	2	1	1	1	1
4	4	2	2	2	2
6	5	4	4	4	3
12	6	5	5	5	4
14	12	6	6	6	5
21	14	11	11	11	6
	16	12	12	12	11
	19	14	13	13	12
	21	16	14	14	13
		19	16	16	14
		21	17	17	16
		23	19	19	17
			20	20	19
			21	21	20
			23	23	21
					23

Table B.14. Parameter estimates after each iteration of MOO design for IG3

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
6.7404	0.4780	3.5917	4.6465	4.8313	4.8846
6.1754	8.1204	9.0180	9.7727	9.7874	9.7825
20.0000	15.8745	15.8745	9.1243	9.6346	9.6676
6.5910	6.7310	7.3259	7.8725	7.9380	7.9379
17.2081	14.5159	7.2785	9.7873	9.8637	9.8597
8.0111	9.3588	9.5121	9.8863	9.8814	9.8876
0.7473	1.1330	0.0788	0.0294	0.0211	0.0252
1.2168	4.2593	4.2673	0.2663	0.0989	0.2687
0.0951	0.0004	0.1263	0.0090	0.0001	0.0782
10.3338	15.8127	0.1402	6.9310	7.3644	7.3903
0.0022	0.0256	0.6307	0.8188	0.9511	0.9511
-1.6113	-0.5432	-1.3871	-1.2080	-0.9334	-0.9397
4.0000	3.9994	3.9996	2.2248	2.1290	2.1117
-0.1815	-1.5633	-1.1076	-0.9303	-0.9341	-0.9358
0.0019	0.0000	0.0000	0.0000	0.0000	0.0000
-1.9999	-0.6658	-0.9262	-0.9655	-0.9876	-0.9876
2.7005	2.2892	2.2046	2.2003	2.2003	2.1139
2.4749	0.0000	0.0000	0.1136	0.0314	0.0116
3.9989	2.4264	2.2367	2.6153	2.1992	2.0795
-0.0407	-2.0000	-2.0000	-0.7357	-0.9360	-0.9373
2.5019	2.4724	2.4704	2.2107	2.0376	2.0270
3.3712	0.7426	0.2838	0.0189	0.0189	0.0873
4.0000	4.0000	3.2420	2.3528	1.8530	1.8878

CURRICULUM VITA

SRIDHARAN SRINATH

BLK 706, #07-383
Clementi West Street 2,
Singapore-120706
Email: srinath@nus.edu.sg

Tel: (Mb): (65) 9642 1916
(Of): (65) 6516 7859

OBJECTIVE

To make interesting & useful contributions in the area of (bio)chemical process modeling and to facilitate student learning.

RESEARCH INTERESTS

Mathematical Modeling of Biological Systems, Multivariate Statistical Data Analysis and Optimization, Biochemical Systems Theory, Systems Biology

ACADEMIC PROFILE

August 2007- March 2012:	***Doctor of Philosophy*** Department of Chemical & Biomolecular Engineering National University of Singapore ***Thesis title*:** Model Identification in the Biochemical Systems Theory
July 2005 - June 2007:	***Master of Technology in Advanced Chemical Analysis (GPA: 9.0/10.0)*** Department of Chemistry, Indian Institute of Technology, Roorkee, India ***Thesis title:*** Theoretical Studies of Stability and Reactivity of Lower Fullerenes
August 2001 – May 2005:	***Bachelor of Technology in Chemical Engineering (81.1%)*** Anna University, Chennai, India ***Project Title:*** Hydrodynamic studies in a tapered fluidized bed reactor and design of water treatment plant.

RESEARCH EXPERIENCE

CO_2 capture and sequestration (Feb 2012 - present)

Advisor(s): Prof. Farooq. S & Prof. I.A. Karimi
Institute: National University of Singapore, Singapore (jointly affiliated with NTU, Singapore)

- Process modeling and optimization of a 4-step Pressure Swing Adsorption (PSA) process
- Cost-Energy optimization to achieve desired recovery-productivity of CO_2.

Dynamic BioSystems Group (Aug 2007 – Mar 2012)

Thesis: Model Identification in the Biochemical Systems Theory
Institute: National University of Singapore, Singapore
Advisor(s): Dr. Rudiyanto Gunawan

- Developed methods for parameter identifiability of ODE models, with special focus on biological systems modeled using Biochemical Systems Theory (BST) models. Extended this method to decoupled and lin-log models under the BST framework. These methods were based on multivariate statistics and nonlinear regression.
- Developed a new multi-objective optimization curvature-based optimal experiment design for ODE models
- Integrated the above methods into a useful tool in MATLAB

Theoretical Chemistry Lab (July 2006 – June 2007)

Thesis: Theoretical Studies of Stability and Reactivity of Lower Fullerenes
Institute: Indian Institute of Technology Roorkee, India
Advisor(s): Dr. P. P. Thankachan

- Studied the structure and properties of the lower fullerenes – C_{20} and C_{24}.
- Also studied the dynamics of their hydrogenation.

TEACHING EXPERIENCE

Tutor, National University of Singapore
CN 2116 Chemical Kinetics and Reactor Design (Sem II, 2009-2010)
IT 1005 Introduction to Computing with MATLAB (Sem I, 2010-2011)
Grader, National University of Singapore
CN 3124 Particulate Technology (Sem II 2007-2008; Sem II 2008-2009)
CN 3421 Process Modelling & Numerical Simulation (Sem II 2007-2008; Sem II 2008-2009)
Supervised 3undergraduate students towards their research project, NUS

INTERNSHIP EXPERIENCE

- Implant training at Lux flavours, Chennai and at Vasudha Pharma Chem Ltd., Hyderabad.
- Worked on selected Analytical Techniques and got familiarized with their principle and operation at **Dr.Reddy's Labortories Ltd**, Hyderabad, India

RESEARCH OUTCOMES

Journal publications:

1. Srinath S and Gunawan R. *Parameter identifiability of power-law biochemical system models.* J Biotechnol 2010 Sep 1; 149(3) 132-40.
2. Srinath S and Gunawan R. *Multiobjective Optimization of Experiments: Curvature and Fisher Information Matrix*, AiChe, *In Review*
3. Srinath S and Gunawan R. *Iterative Model Identification of BST models*, BMC Systems Biology, *In Preparation*

Conference Proceedings:

1. Srinath S, Gunawan R. **(2011)** *Model-based Design of Experiment for Kinetic Parameter Identification: Beyond the Fisher Information Matrix*, 12th International Conference on Molecular Systems Biology (ICMSB), Lleida, Spain, May 8-12
2. Srinath S, Gunawan R. **(2010)** Parameter Identifiability of Metabolic Network Models, In *Satellite Conference of the International Congress of Mathematics*, Hyderabad, India, Aug 16-18
3. Srinath S, Yuan Z, Gunawan R. **(2010)** *Identifiability Analysis of Decoupled Power-Law Models*, 5th International Symposium on Design, Operation and Control of Chemical Processes (PSE Asia), Singapore, July 25-28

4. Srinath S, Gunawan R. **(2010)** *Parameter Identifiability in Kinetic Modeling of Metabolic Pathways*, Poster presented at the Metabolic Engineering Conference VIII, Jeju Island, South Korea, Jun 13 - 17
5. Srinath S, Gunawan R. **(2009)** Identifiability Analysis of Metabolic Networks, 11th International Conference on Molecular Systems Biology (ICMSB), Shanghai, China, June 21-25

TECHNICAL SKILLS

- Extensive knowledge and familiarity with MATLAB
- C/C++ programming and hands on knowledge in COMSOL Multiphysics, GAMS and HySys

ADDITIONAL TRAINING

- Communication skills workshop, NUS, March, 2008
- Training Course for Teaching Assistants, December, 2009

HONORS & ACTIVITIES

- Department of Chemistry topper in M.Tech in IIT Roorkee (July 2005- July 2007)
- NUS Research Fellowship – Aug 2007 to Aug 2011
- Actively participated in the organizing committee of PSE Asia 2010
- Treasurer, Graduate Students' Association, National University of Singapore, Singapore (2008 – 2009) and coordinated in organizing "Chembiotech '08", a regional conference in NUS.

PERSONAL PROFILE

Nationality	:	Indian
Date of Birth	:	10th May 1984
Gender	:	Male
Marital Status	:	Married
Alternate Id	:	srinath.svce@gmail.com
Languages Known	:	English, Hindi, Tamil and Telugu

REFERENCES

Dr. Lakshminarayanan S.
Department of Chemical & Biomolecular Engineering,
National University of Singapore,
Singapore. Tel: (65) 6516 8484
Email: chels@nus.edu.sg

Prof. I.A. Karimi
Department of Chemical & Biomolecular Engineering,
National University of Singapore
Singapore- Tel: (65) 6516 6359
Email: cheiak@nus.edu.sg

DECLARATION

I hereby declare that all the details furnished above are true to the best of my knowledge.

PLACE: Singapore
DATE: October 11, 2012 **SRIDHARAN SRINATH**